IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sumia Tahir
Oct 13, 2022

Outline

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

Executive Summary

- We extracted data from Space X’s REST API and also from the Falcon 9 historical launch records on wikipedia to investigate the success rate of SpaceX’s landing data and launch successes.
- We found that the launch success increased with flight number.
- We also found that launch success varies with booster version, payload mass and orbit type.
- Our machine learning model was able to predict launch success rate with 83% accuracy.
- The landing success rate of the first stage was higher on ground pad than on drone ship.

Introduction

- During the current space age, companies such as Virgin Galactic, Rocket Lab and Blue Origin are in a race to make space travel affordable for everyone.

- SpaceX is a leader in this field. SpaceX claims that its Falcon 9 rocket launches cost 62 million dollars while the cost of launches by other competitors is more than 165 million dollars per launch. Since SpaceX can reuse the first stage of its rocket, this accounts for most of the savings.

- Working at a new rocket company, Space Y, our goal is to predict the price of each launch. We will do this by gathering data on SpaceX's launches to determine the landing success of their first stage. If the first stage can be reused, then the price of the launch can be approximated.

- If Space Y wants to be a solid competitor in space travel, then we will need to keep the cost of our launches close to those of SpaceX.

Section 1

Methodology

Methodology

- Data collection methodology
- Data wrangling
- EDA using visualization and SQL
- Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis using classification models

Data Collection

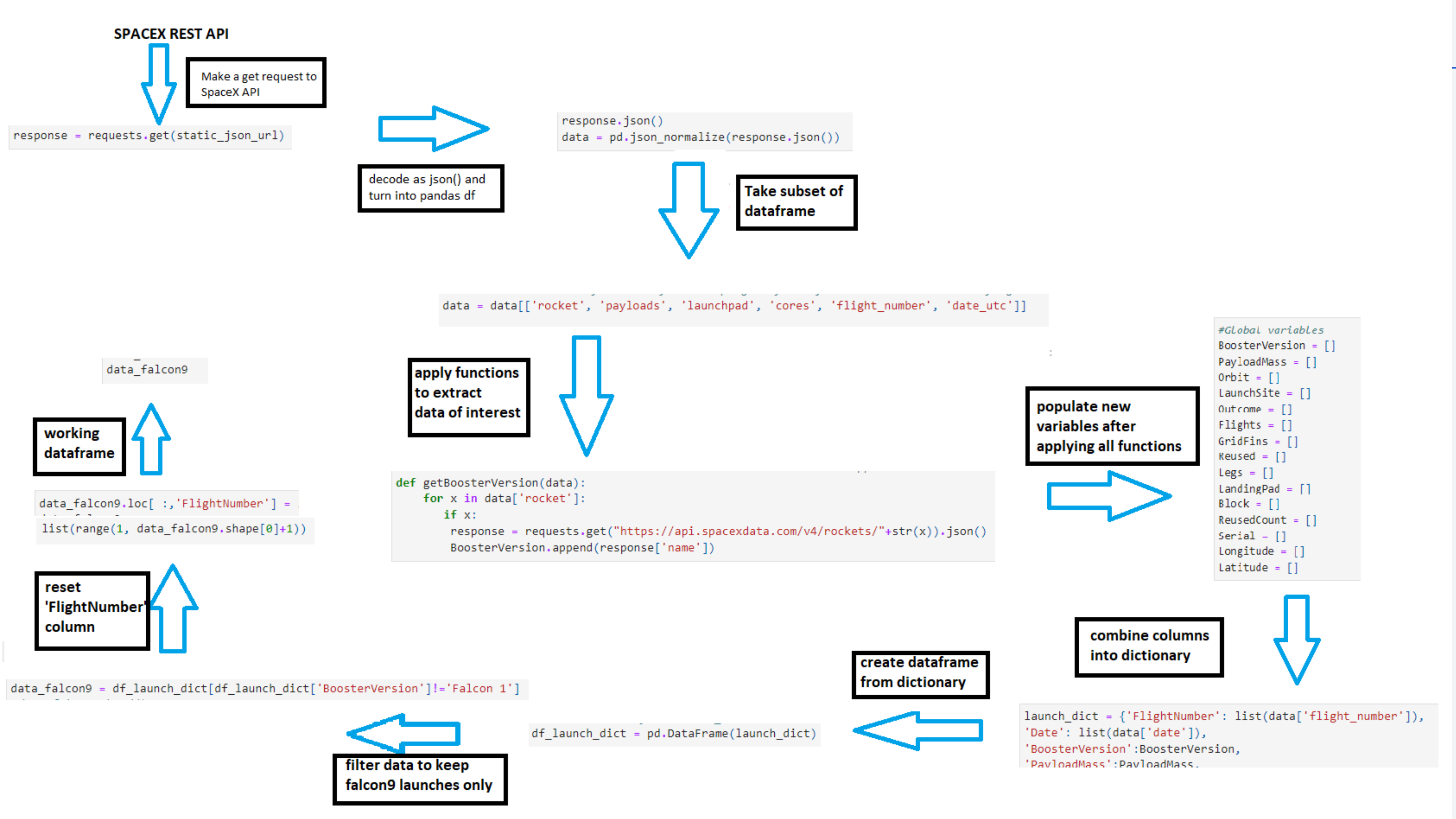
The original data was obtained from the SPACEX REST API by making a Get request.

A modified dataset was used from the following url:

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

The data was filtered to only include records from the Falcon 9 rocket.

Data Collection – SpaceX API



[Notebook: Capstone-Project/Data Collection API.ipynb at main · stahir21/Capstone-Project (github.com)]

Data Collection – SpaceX API

Dataframe was checked for missing values:

```
data_falcon9
```

check for missing values

```
data_falcon9.isnull().sum()
```

missing values in 2 features

```
PayloadMass        5
LandingPad        26
```

** ‘LandingPad’ column will retain *None* values to account for when landing pads were not used

calculate mean for payload mass

```
pay_mean = data_falcon9['PayloadMass'].mean(axis=0)
```

replace missing value with mean

```
data_falcon9['PayloadMass'].replace(np.nan, pay_mean, inplace=True)
```


[Notebook: Capstone-Project/Data Collection API.ipynb at main · stahir21/Capstone-Project (github.com)]

Data Collection

Falcon 9 historical launch records were webscraped from a Wikipedia page titled "List of Falcon 9 and Falcon Heavy launches"

https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Data Collection - Webscraping

```
static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"
```

request HTML and get response object, create BeautifulSoup object

```
response = requests.get(static_url).text
soup = BeautifulSoup(response, 'html.parser')
```

find all tables on wiki page

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

extract launch table

```
first_launch_table = html_tables[2]
```

get all column names

```
column_names = []
headers = first_launch_table.find_all('th')

for header in headers:
    name = extract_column_from_header(header)
    if name is not None and len(name)>0:
        column_names.append(name)
```

```
['Flight No.', 'Date and time ( )', 'Launch site', 'Payload', 'Payload mass', 'Orbit', 'Customer', 'Launch outcome']
```

create empty dictionary with keys from the extracted column names

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

del/add columns

```
del launch_dict['Date and time ( )']
```

```
# Added some new columns
launch_dict['Version Booster']=[]
launch_dict['Booster landing']=[]
launch_dict['Date']=[]
launch_dict['Time']=[]
```

extract data from tables to fill dictionary

```
for table_number,table in enumerate(soup.find_all('table',"wikitable plainrowheaders collapsible")):
    for rows in table.find_all("tr"):
        if rows.th:
            if rows.th.string:
                flight_number=rows.th.string.strip()
                flag=flight_number.isdigit()
```

create dataframe from dictionary

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


[Notebook: Capstone-Project/jupyter-labs-webscraping.ipynb at main · stahir21/Capstone-Project (github.com)]

Data Wrangling

- Objective: perform EDA to find patterns in the data and determine labels for training supervised models
- Convert outcomes into training labels where 1 means the booster landed successfully, and 0 means it was unsuccessful

Data Wrangling

```
df=pd.read_csv("https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/dataset_part_1.csv")
```

find number of launches on each site

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

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

calculate number and occurrence of each orbit

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

```
GTO      27
ISS      21
VLEO     14
PO        9
LEO       7
SSO       5
MEO       3
ES-L1     1
HEO       1
SO        1
GEO       1
```

find number of landing outcomes and assign to new variable

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

```
True ASDS      41
None None      19
True RTLS      14
False ASDS      6
True Ocean      5
False Ocean     2
None ASDS       2
False RTLS      1
```

get keys for landing_outcomes

```
for i,outcome in enumerate(landing_outcomes.keys()):
    print(i,outcome)
```

```
0 True ASDS
1 None None
2 True RTLS
3 False ASDS
4 True Ocean
5 False Ocean
6 None ASDS
7 False RTLS
```

select outcomes where landing failed

```
bad_outcomes=set(landing_outcomes.keys()[[1,3,5,6,7]])
bad_outcomes
```

```
{'False ASDS', 'False Ocean', 'False RTLS', 'None ASDS', 'None None'}
```

create new variable landing_class where 1=success and 0=failure

```
landing_class = []
for i in range(90):
    if df['Outcome'][i] in bad_outcomes:
        landing_class.append(0)
    else:
        landing_class.append(1)
```

```
df['Class']=landing_class
```


[Notebook: Capstone-Project/labs-jupyter-spacex-Data wrangling.ipynb at main · stahir21/Capstone-Project (github.com)]

EDA with SQL

- Names of all the unique launch sites
- 5 records where launch sites begin with the string 'CCA‘
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Date of the first successful landing outcome in ground pad
- Names of boosters which had successful landings on drone ship with payload mass between 4000 & 6000 kg
- Total number of successful and failure mission outcomes
- Names of the booster_versions that have carried the maximum payload mass
- Month names, booster versions, launch_site for failed landings on drone ship during the year 2015
- Count of landing_outcomes between 04-06-2010 and 20-03-2017 in descending order

Notebook: Capstone-Project/jupyter-labs-eda-sql-coursera_sqllite.ipynb at main · stahir21/Capstone-Project (github.com)

EDA with Data Visualization

The following plots were made to explore the relationships between the different features and to determine their effect on landing outcome:

- Scatterplot of Flight Number vs Launch Site overlaid with Class (success/fail)
- Scatterplot of Payload Mass vs Launch Site overlaid with Class
- Bar chart of Success Rate vs Orbit Type
- Scatterplot of Orbit Type vs Flight Number overlaid with Class
- Scatterplot of Orbit Type vs Payload Mass overlaid with Class
- Line plot of average Launch Success rate vs Year

Notebook: Capstone-Project/jupyter-labs-eda-dataviz.ipynb at main · stahir21/Capstone-Project (github.com)

Interactive Map with Folium

Along with some other factors such as payload mass and orbit type, the success of a launch also depends on the initial position of rocket trajectories. Therefore, the location of a launch site and its proximities affect the success of a launch. We explored the surroundings of existing launch sites to determine factors that can help select optimal locations for building new launch sites.

A Folium map was created with the following objects:

- Circles added to folium map to indicate launch sites
- Markers added to label launch sites
- Green and red marker clusters added for all launch records to indicate success or fail
- Lines added to display distances to nearby railroads, highways, cities and coastlines

IBM-Capstone-Project/lab_jupyter_launch_site_location.ipynb at main · stahir21/IBM-Capstone-Project (github.com)

Dashboard with Plotly Dash

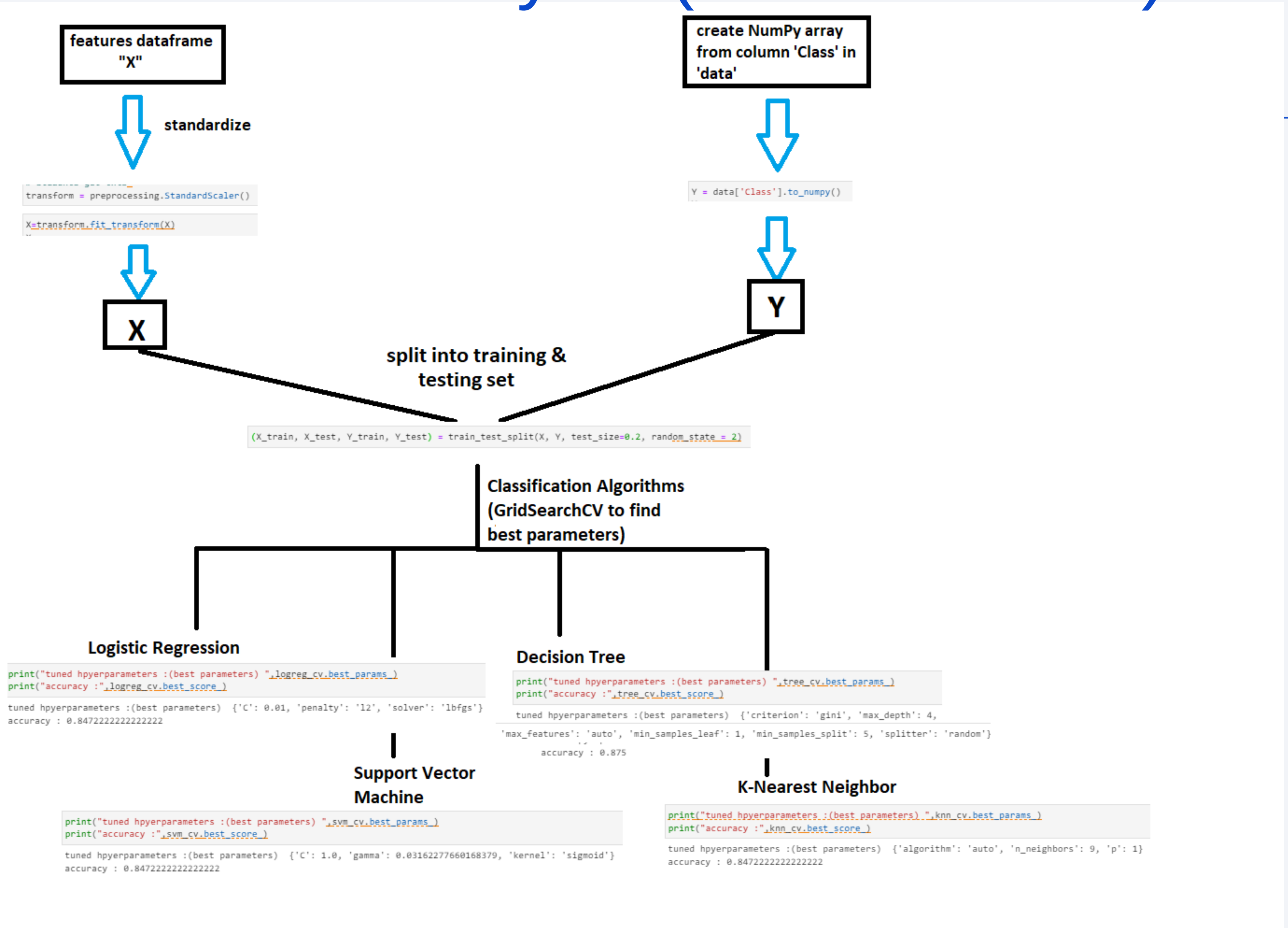
The dashboard includes the following:

- Launch site drop-down menu allows selection of all launch sites or of one specified site
- Pie chart shows the counts of successful launches for all launch sites or for the selected launch site
- Range slider allows adjustment of the payload variable to see how different payload masses affect mission outcome
- Scatterplot shows relationship between payload mass and the launch outcome, with color-coding for different booster versions. A scatterplot can be generated for individual launch sites.

Notebook:

IBM-Capstone-Project/spacex_dash_app.py at main · stahir21/IBM-Capstone-Project (github.com)

Predictive Analysis (Classification)



[Notebook: https://github.com/stahir21/IBM-Capstone-Project/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5%20(1).ipynb](https://github.com/stahir21/IBM-Capstone-Project/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5%20(1).ipynb)

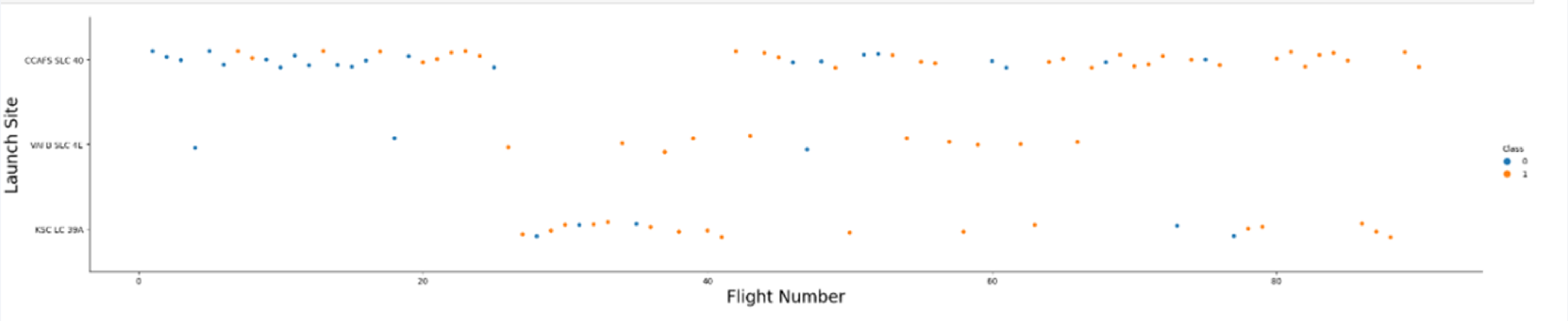
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

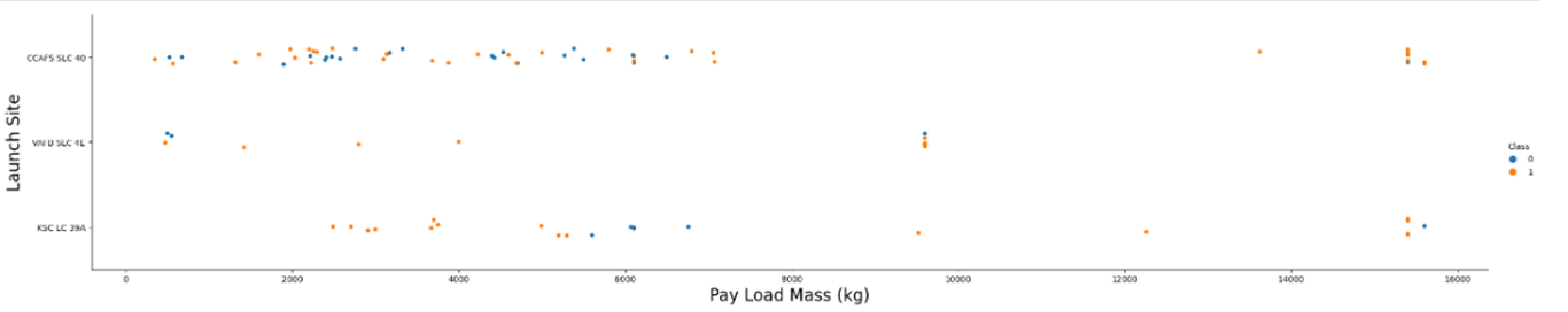


This plot shows the relationship between Launch Site and Flight Number.

The most flights were launched from the Cape Canaveral Launch Complex and the least number of flights were from Vandenberg Launch Complex.

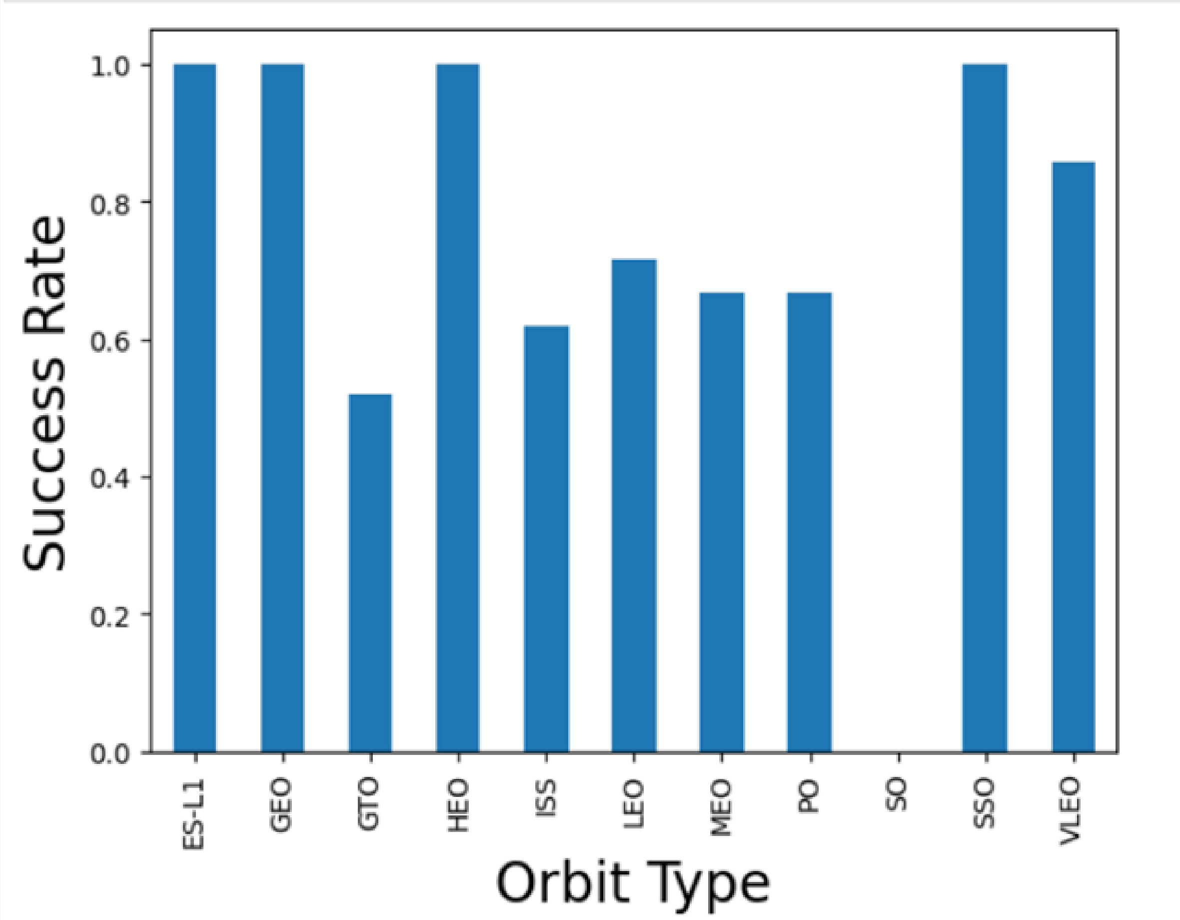
The success rate was higher with increasing flight number.

Payload Mass vs. Launch Site



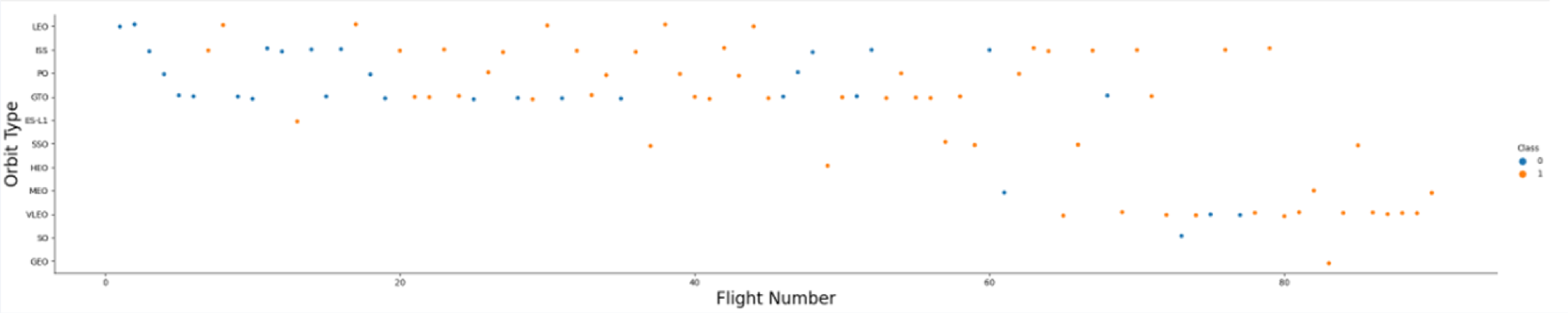
The Vandenberg Launch Complex did not launch any boosters with payloads above 10,000 kg. Most of the launches took place at Cape Canaveral.

Success Rate vs. Orbit Type



- The orbits ES-L1, GEO, HEO and SSO had success rates close to 100%. The GTO orbit was the least successful with rate of around 50%.

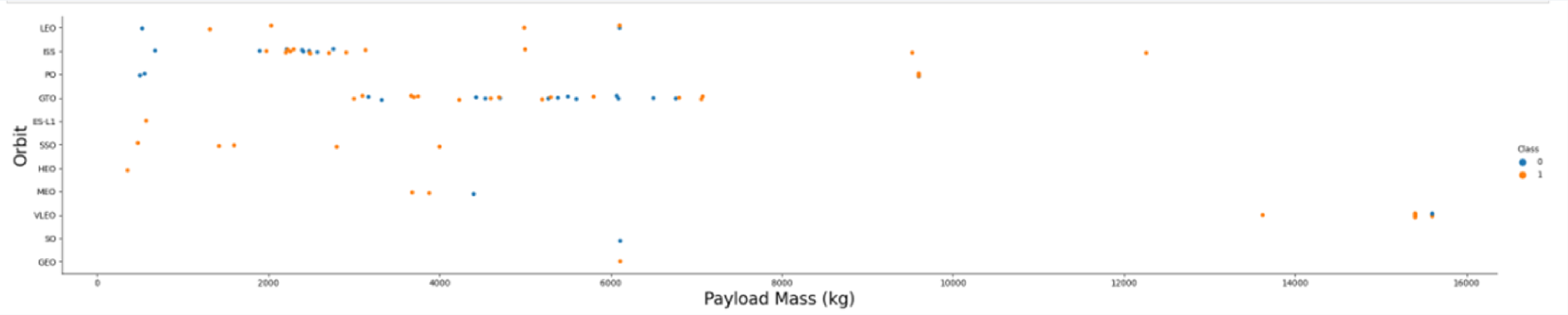
Flight Number vs. Orbit Type



For the LEO orbit, there are more successful outcomes with increasing flight number.

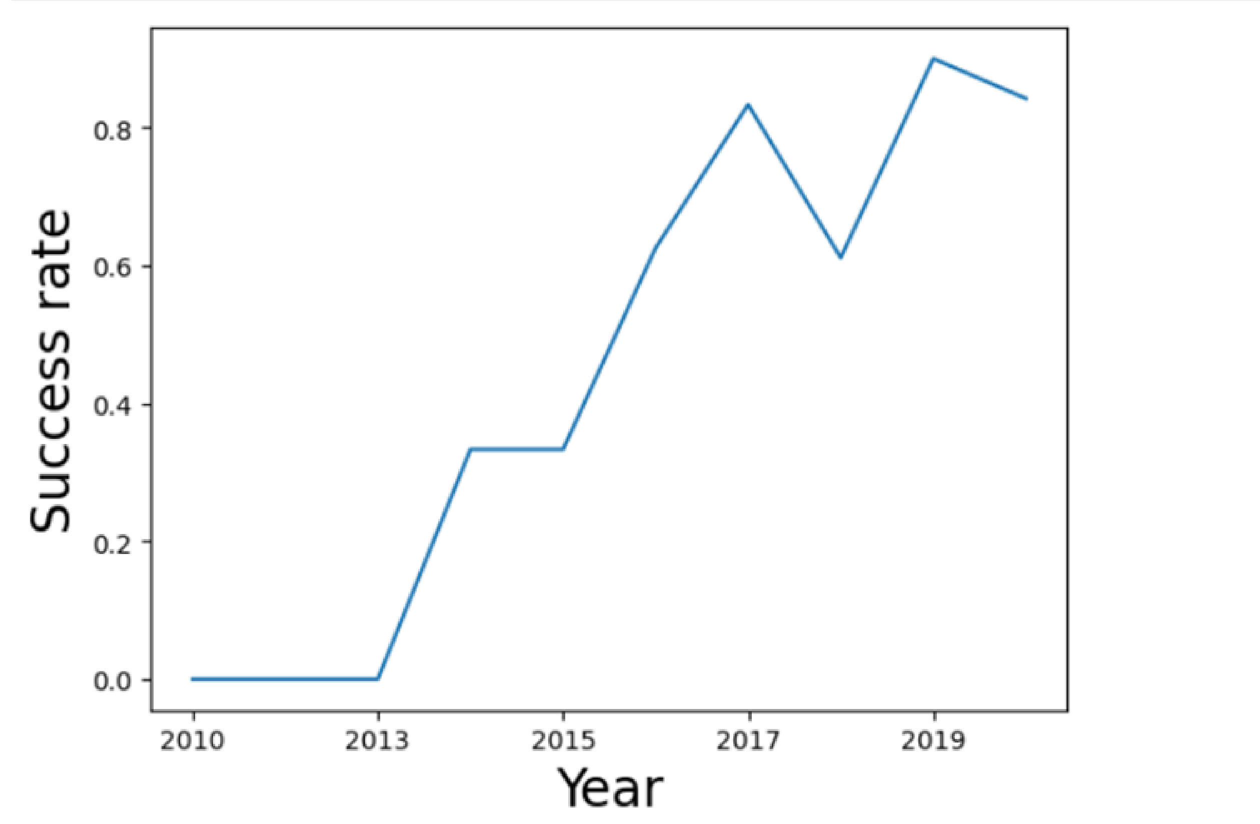
Some orbits that we observed had 100% success rate (see previous bar graph) had only 1 flight (eg. GEO, HEO, ES-L1).

Payload vs. Orbit Type



For the heavier payloads, the orbits LEO, ISS, and VLEO had more successes.

Launch Success Yearly Trend



The launch success rate has been increasing since 2013.

All Launch Site Names

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

Cape Canaveral Launch Complex 40
Vandenberg Space Launch Complex 4 East
Kennedy Space Center Launch Complex 39A
Cape Canaveral Space Launch Complex 40

Note: CCAFS LC-40 and CCAFS SLC-40 are the same launch complex.
The name was changed from LC-40 to SCL-40 in 1997.

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

5 records are displayed where the launch site name begins with ‘CCA’.
The launch site is Cape Canaveral Launch Complex 40. The launches took place between 2010-2013.
In 3 of the 5 displayed cases, the landing of the booster was not attempted.
In the other 2 cases, the landing was a failure.

Total Payload Mass

SUM(PAYLOAD_MASS__KG_)
48213

The total payload mass carried by boosters launched by NASA (CRS) is 48,213 Kg.

Average Payload Mass by F9 v1.1

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

The average payload mass carried by booster version 1.1 of the Falcon 9 rocket was **2534.7kg.**

First Successful Ground Landing Date

Date	Landing _Outcome
22-12-2015	Success (ground pad)

The first successful landing of the Falcon 9 rocket on ground pad was on Dec 22, 2015.

Successful Drone Ship Landing with Payload between 4000 and 6000

Booster_Version	Landing _Outcome	PAYLOAD_MASS__KG_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

The booster versions listed above were able to land successfully on drone ships while carrying a payload mass between 4000 and 6000 kg.

Total Number of Successful and Failure Mission Outcomes

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

The total number of successful mission outcomes was 100 and there was 1 failure mission outcome.

Boosters Carried Maximum Payload

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

The maximum payload mass carried was 15,600 kg.
The booster versions that carried the maximum load are listed in the table.

2015 Launch Records

Month	Year	Booster_Version	Launch_Site	Landing _Outcome
01	2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

In January and April of 2015, there were 2 failed landings on drone ships by boosters B1012 and B1015.

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

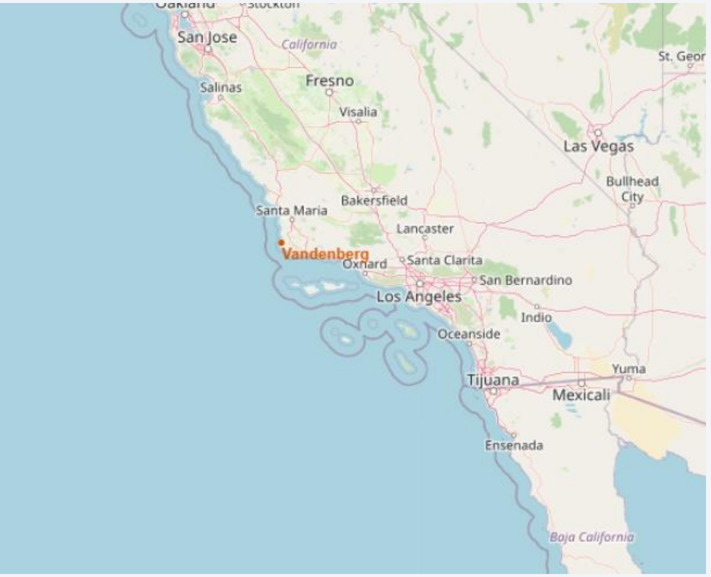
Landing _Outcome	COUNT("Landing_Outcome")
Success (drone ship)	12
No attempt	12
Success (ground pad)	8
Failure (drone ship)	5
Controlled (ocean)	4
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

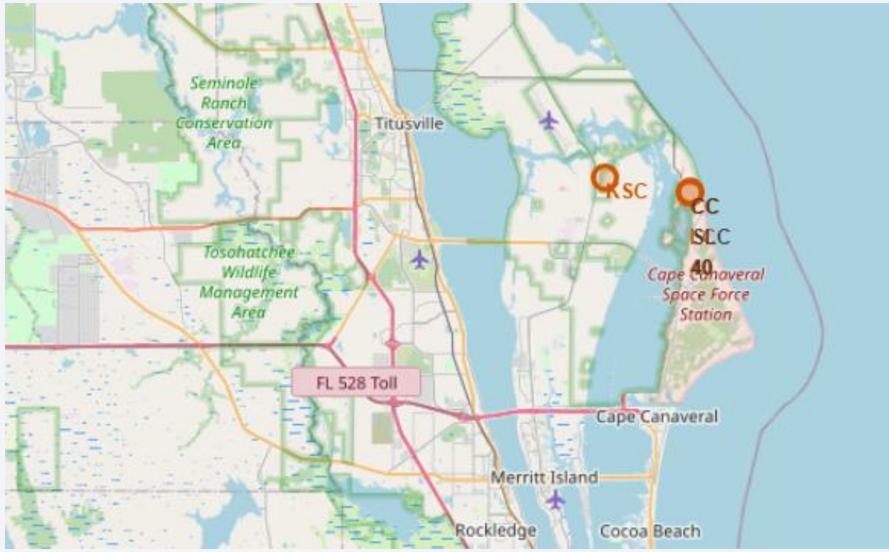
The most common landing outcomes between 2010-06-04 and 2017-03-20 were successful landing on drone ship, no attempt, successful landing on ground pad and failed landing on drone ship.

Section 3

Launch Sites Proximities Analysis

Folium map – Launch Sites

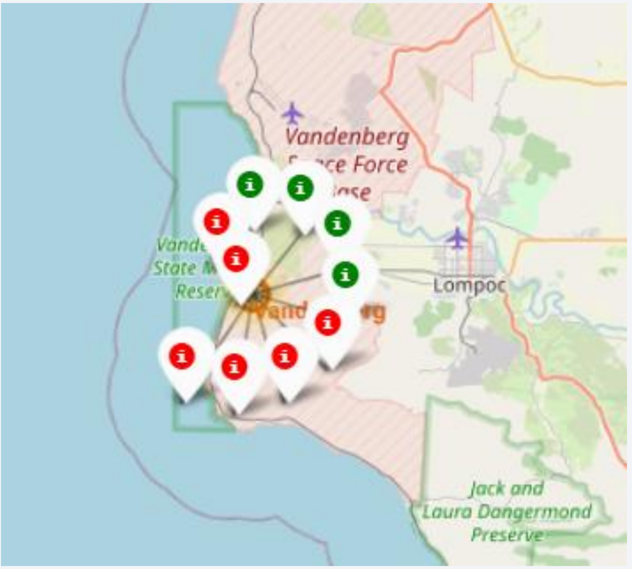




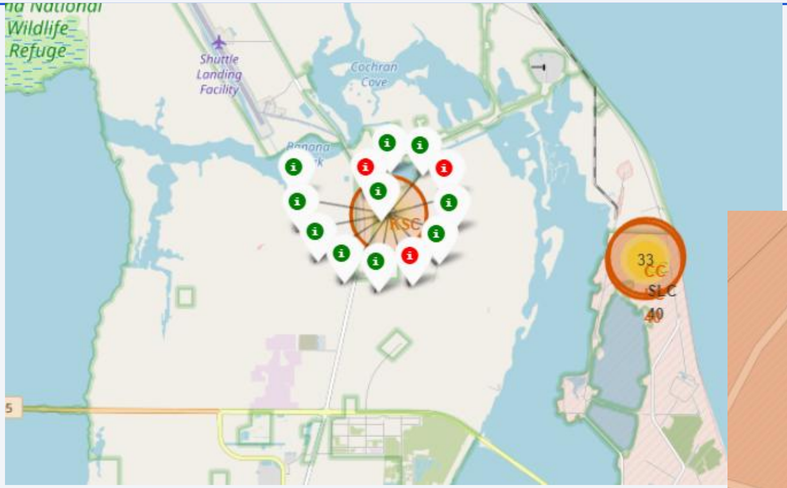
Current Space Launch sites are displayed:
Cape Canaveral Launch complex, Kennedy space center, Vandenberg Space Complex

The launch sites are located in close proximity to coastlines and are located in the southern most states of the USA so that they are close to the equator.

Folium map: Marker clusters for launch success/fail

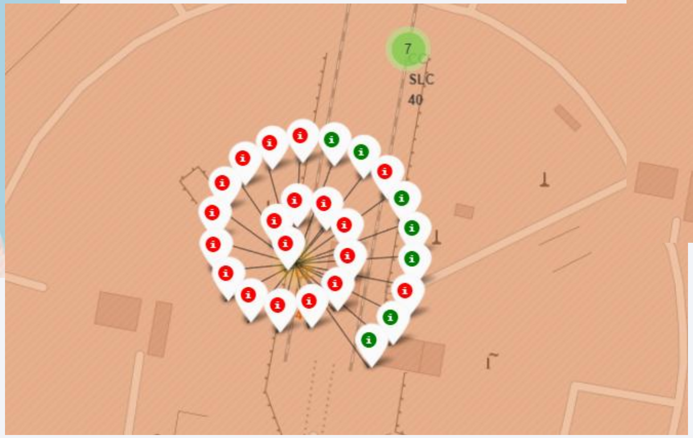


Vandenberg



Kennedy Space Center

Cape Canaveral SLC 40

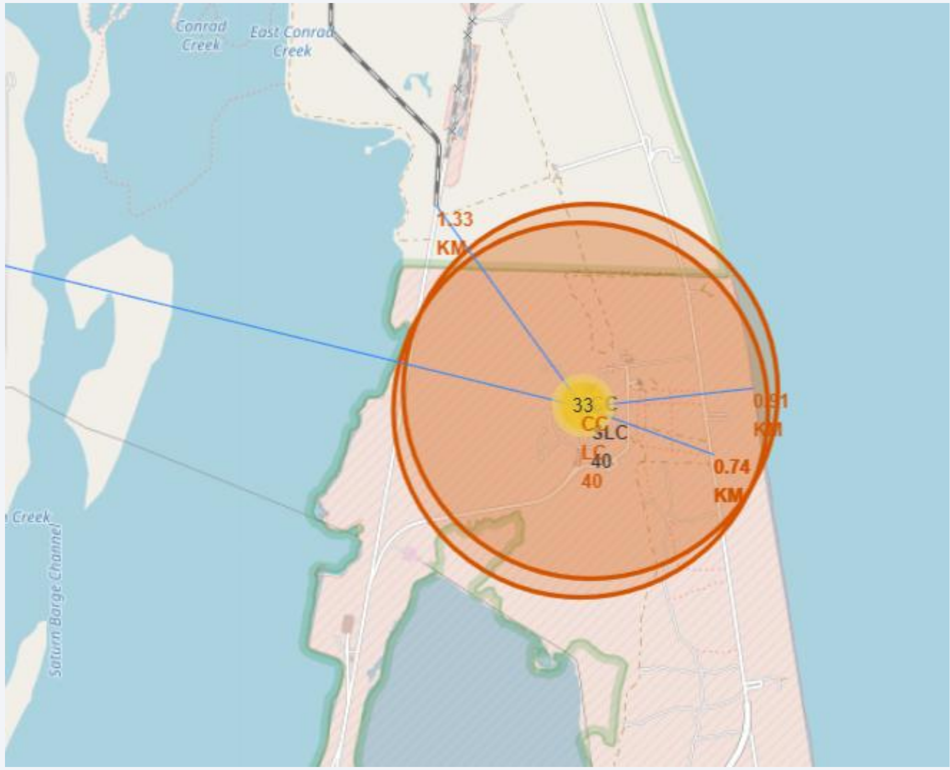


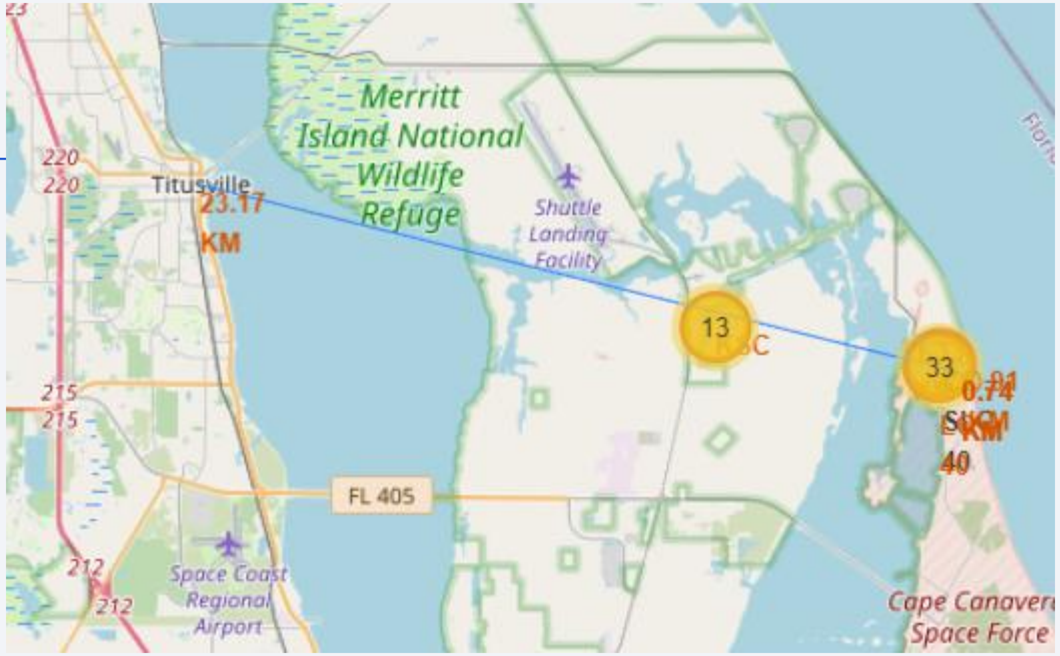
Cape Canaveral LC 40

Each marker represents a launch record. Green markers indicate successful launches while red markers indicate failed launches.

Kennedy Space Center has the highest rate of successful launches while Cape Canaveral (LC-40) has the lowest rate.

Folium map: distances to proximities



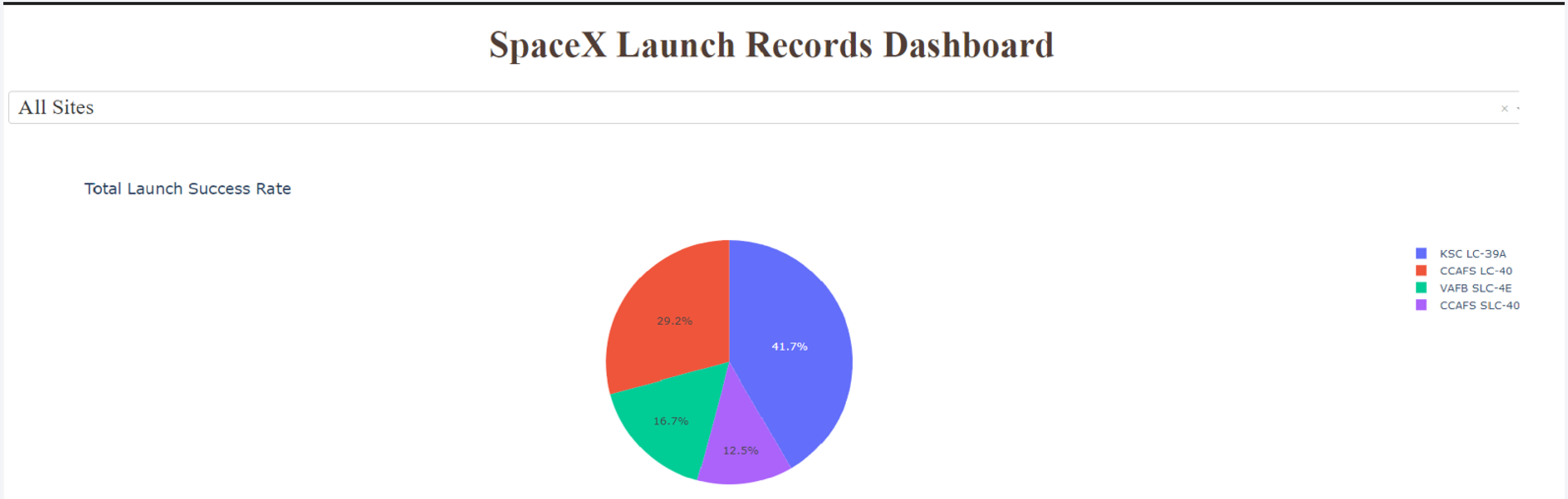


The distances from Cape Canaveral (LC-40) to nearby proximities are displayed.
The nearest city, Titusville, is located 23 km away.
The railway line, highway and coastline are all located very close to the launch site, around a 1 km radius.

Section 4

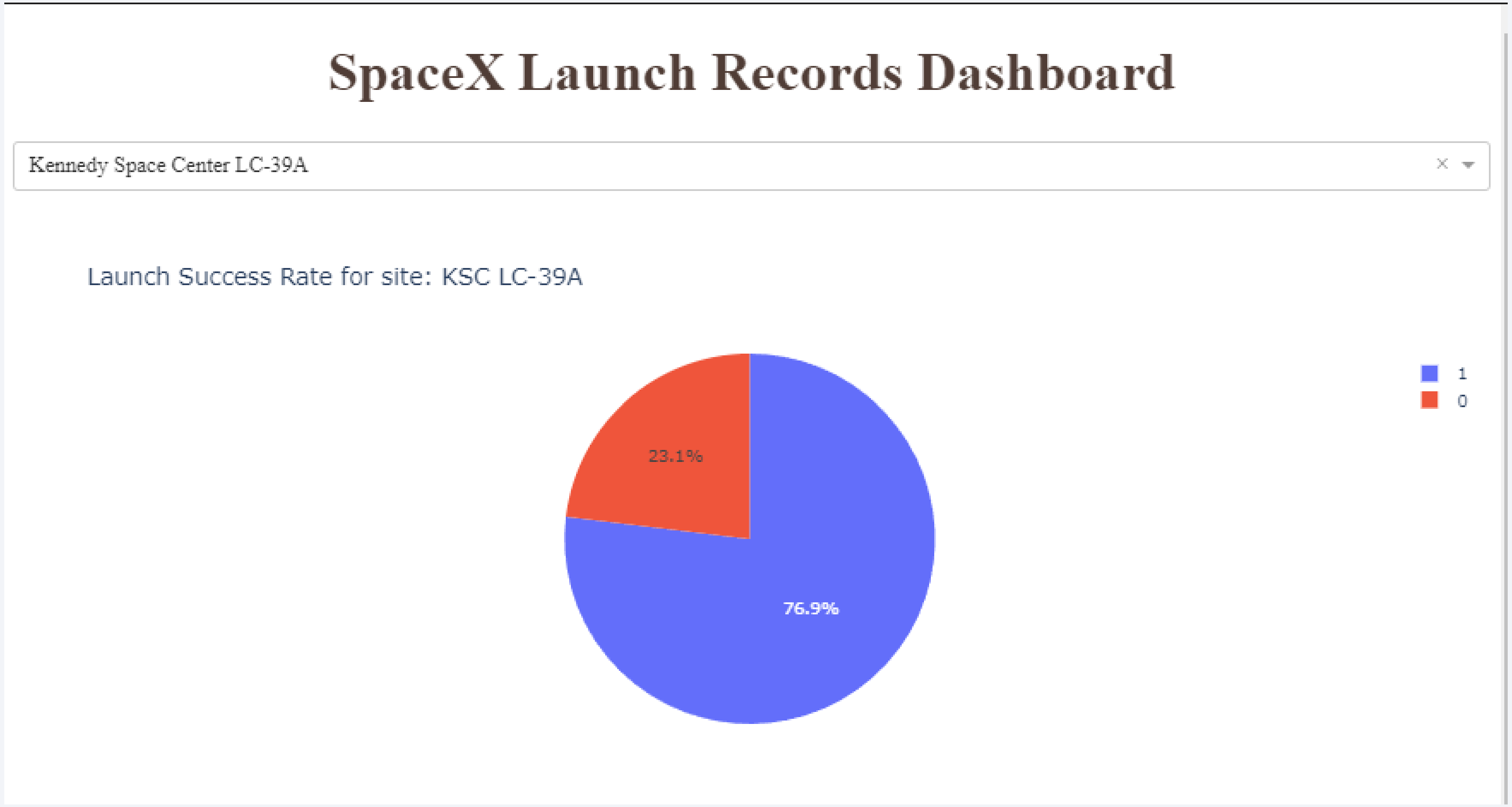
Build a Dashboard with Plotly Dash

Dashboard – Launch success rate (all sites)



The pie chart shows that 41% of the successful launches were from Kennedy Space Center.

Dashboard: Launch Success Rate of KSC



Kennedy Space Center had a success launch rate of 77%.

Dashboard: payload and success rate

Across the full payload range, the FT Booster version had the most successes while the v1.1 booster version had the most failures.

The launch success rate was low (3/11 = 27%) for payloads between 5000-10000 kg.

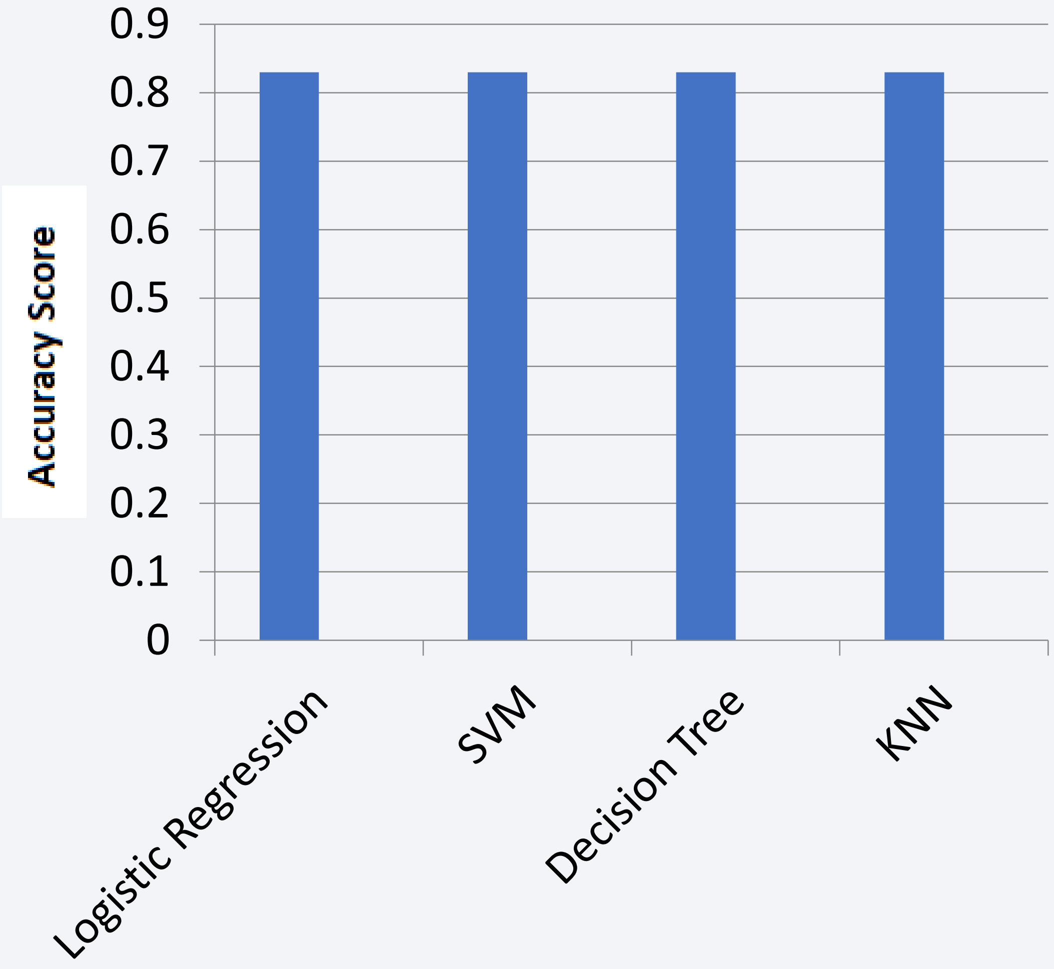
The launch success rate was 54% (15/28) for payloads between 2000-5000 kg.

Note: when the payload slider is adjusted, the legend colors also change so the same boosters may be displayed in different colors in different plots.

Section 5

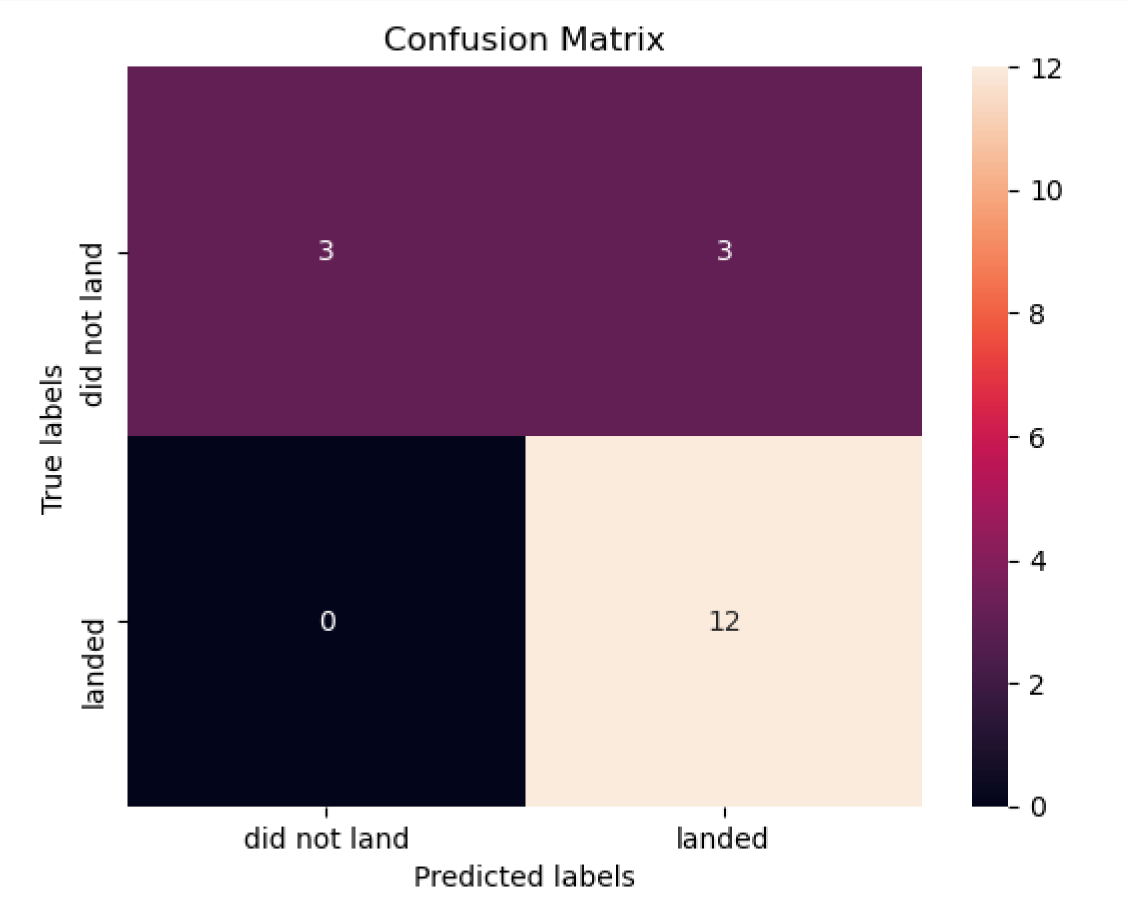
Predictive Analysis (Classification)

Classification Accuracy



All classification algorithms performed equally well.
The accuracy score on the test set was 0.83 for all methods.

Confusion Matrix



The same confusion matrix was generated by all classification algorithms.
However, there is an issue with False positives (top-right corner).
In 3 cases, the rocket did not land successfully but the models predicted that the landings were successful.

Conclusions

- Space Y would like to enter the space race and be a solid competitor with other space companies. Currently, SpaceX leads the way because they advertise that their launches only cost 62 million dollars per launch as compared to 165 million dollars by other companies. This cost reduction is primarily due to their reuse of the first stage of the rocket.
- We analyzed SpaceX's launch records to predict their success rate. This will allow us to accurately determine their cost of launches so that our company can aim to reduce our cost of launches accordingly. We also investigated the factors that led to successful launches and landing outcomes so that we can implement those factors in our missions.
- We found that the launch success rate increased with more flights. The bar chart showed an increasing trend from 2013. This is expected because with every new flight, improvements and modifications would have been made.
- Exploring the landing outcomes, we found that the landing success rate on ground pad (8 successes, 0 failures) was higher than the rate on drone ship (12 successes, 5 failures). Therefore, ground pad landings should be preferred.

Conclusions

- With regards to launch success rate, Kennedy Space Center had the highest rate of successful launches compared to the other sites. We also found that launch sites are located near coastlines and the equator, with nearby railways and railroads. However, they are located further away from cities. Therefore, Space Y can build their launch sites at locations that have similar proximities.
- The launch success rate was higher for boosters carrying a payload mass between 2000-5000kg than for payload masses of 5000-10000kg (see dashboard payload slider).
- All of the classification algorithms performed equally well. They were able to predict successful launches with an accuracy score of 83%. Therefore, Space Y has an accurate idea of what factors will result in a successful launch.

Appendix

<u>**Python Code for Plots:**</u>

- **Scatterplot of Launch Site vs Flight Number**

```
sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number", fontsize=20)
plt.ylabel("Launch Site", fontsize=20)
plt.show()
```

- **Scatterplot of Payload Mass vs Launch Site**

```
sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df, aspect=5)
plt.xlabel("Pay Load Mass (kg)", fontsize=20)
plt.ylabel("Launch Site", fontsize=20)
plt.show()
```

- **Bar Chart of Success Rate for Orbit Types**

```
df.groupby(['Orbit']).mean()['Class'].plot(kind='bar')
plt.xlabel("Orbit Type",fontsize=20)
plt.ylabel("Success Rate",fontsize=20)
plt.show()
```

- **Scatterplot of Flight Number vs Orbit Type**

```
sns.catplot(y="Orbit", x="FlightNumber", hue="Class", data=df, aspect=5)
plt.xlabel("Flight Number", fontsize=20)
plt.ylabel("Orbit Type", fontsize=20)
plt.show()
```

Appendix

<u>Python Code for Plots Cont'd:</u>

- **Scatterplot of Orbit Type vs Payload Mass**

```
sns.catplot(y="Orbit", x="PayloadMass", hue="Class", data=df, aspect=5)
plt.xlabel("Payload Mass (kg)", fontsize=20)
plt.ylabel("Orbit", fontsize=20)
plt.show()
```

- **Launch success yearly trend**

```
df.groupby(['Year']).mean()['Class'].plot(kind='line')
plt.xlabel("Year",fontsize=20)
plt.ylabel("Success rate",fontsize=20)
plt.show()
```

Appendix

<u>**SQL queries:**</u>

- **All launch site names**

```
%%sql
SELECT distinct(Launch_Site)
FROM SPACEXTBL
```

- **Launch Site Names begin with CCA**

```
%%sql
SELECT *
FROM SPACEXTBL
WHERE Launch_Site LIKE 'CCA%'
LIMIT 5
```

- **Total payload mass**

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE Customer LIKE '%CRS%'
```

- **Average payload mass by F9 v1.1**

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE Booster_Version LIKE '%F9 v1.1%'
```

Appendix

SQL queries cont'd:

- **First successful landing on ground pad**

```
%%sql
SELECT Date, "Landing _Outcome"
FROM SPACEXTBL
WHERE "Landing _Outcome" == "Success (ground pad)"
LIMIT 1
```

- **Successful landing on drone ship with payload between 4000 and 6000**

```
%%sql
SELECT distinct Booster_Version, "Landing _Outcome", PAYLOAD_MASS__KG_
FROM SPACEXTBL
WHERE ("Landing _Outcome" == 'Success (drone ship)') AND (PAYLOAD_MASS__KG_ > 4000) AND (PAYLOAD_MASS__KG_ < 6000)
```

- **Total number of successful and failure mission outcomes**

```
%%sql
SELECT Mission_Outcome, COUNT(Mission_Outcome)
FROM SPACEXTBL
GROUP BY Mission_Outcome
```

- **Boosters that carried maximum payload mass**

```
%%sql
SELECT DISTINCT Booster_Version, PAYLOAD_MASS__KG_
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ == (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

Appendix

SQL queries cont'd:

- **2015 launch records**

```
%%sql
SELECT substr(Date,4,2) AS Month, substr(Date,7,4) AS Year, Booster_Version, Launch_Site, "Landing _Outcome"
FROM SPACEXTBL
WHERE "Landing _Outcome" == 'Failure (drone ship)' AND substr(Date,7,4)='2015'
```

- **rank of landing outcomes between 2010-06-04 and 2017-03-20**

```
%%sql
SELECT "Landing _Outcome", COUNT("Landing _Outcome")
FROM SPACEXTBL
WHERE substr(Date,7,4) < '2018'
GROUP BY "Landing _Outcome"
ORDER BY COUNT("Landing _Outcome") DESC
```

Appendix

<u>Code for Prediction (example – logistic regression):</u>

GridSearchCV

```
lr_parameters ={"C":[0.01,0.1,1],'penalty':['l2'], 'solver':['lbfgs']}# l1 lasso l2 ridge
lr=LogisticRegression()
logreg_cv = GridSearchCV(lr, lr_parameters, cv=10)
logreg_cv.fit(X_train,Y_train)
```

Accuracy score on the test set

```
logreg_cv.score(X_test, Y_test)
```

```
0.8333333333333334
```

Confusion matrix

```
yhat=logreg_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```

Thank you!