IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Riku Sundell
2024/01/01

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Machine learning was used to identify and predict important variables for predicting SpaceX's success with the Falcon 9 (successful landings)
- It was found that parameters like launch site location, launch orbit and payload mass were key variables determining successful landings
- Three ML models were developed that were able to predict successful landings at 83% accuracy on test data

Introduction

- SpaceX has unique capabilities enabling them to provide launches at $62m while competitors' prices are above $165m
- SpaceX Falcon 9 is the first commercial rocket with reusable first stage (booster)
- Upon successful landing SpaceX can provide the first stage for new flights thus lowering cost
- **We wish to predict successful booster reuse**
- Questions to be examined:
  - **What parameters determine a successful landing?**
  - **What is needed for successful launch operations?**

Section 1

Methodology

Methodology

Executive Summary

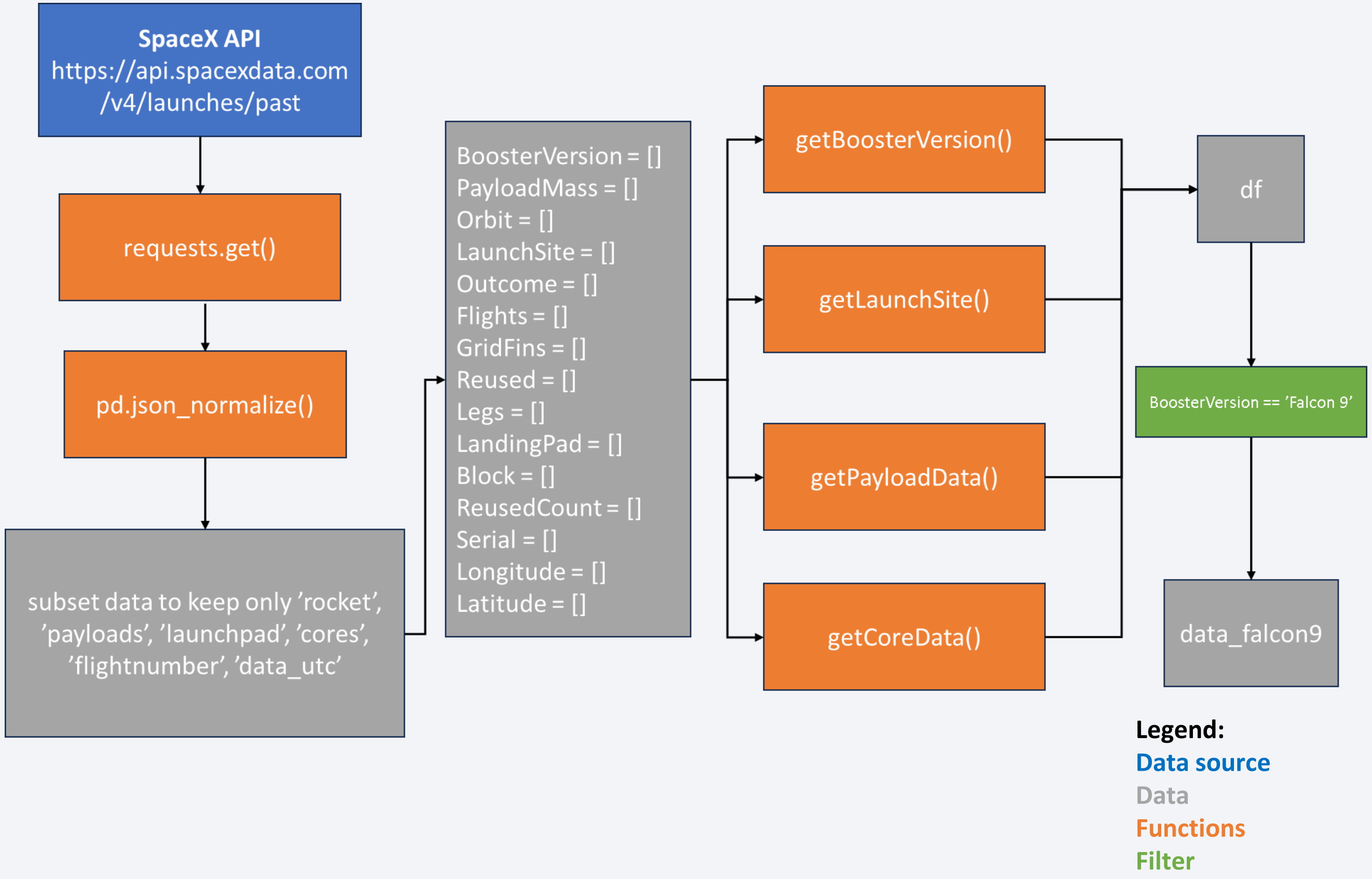
- Launch data was collected from Wikipedia and from SpaceX API, and missing data was parsed or removed
- Exploratory data analysis (EDA) was performed using visualization with Pandas and Matplotlib, and SQL
- Visual analytics was performed using Folium and Plotly Dash
- Predictive analysis was performed using classification models with sklearn:
    - Build
    - Tune
    - Evaluation of models

Data Collection

- Data was collected from SpaceX API and Wikipedia
    - Data for individual launches were requested and wrangled into a data frame with SpaceX Falcon 9 launches from SpaceX API
    - Missing values for payload mass were replaced with the calculated average payload mass
    - Using BeautifulSoup, the Wikipedia page for SpaceX Falcon 9 was web scraped to extract launch records
    - This was parsed and converted into a Pandas data frame
- These steps are presented in more detail in the following slides

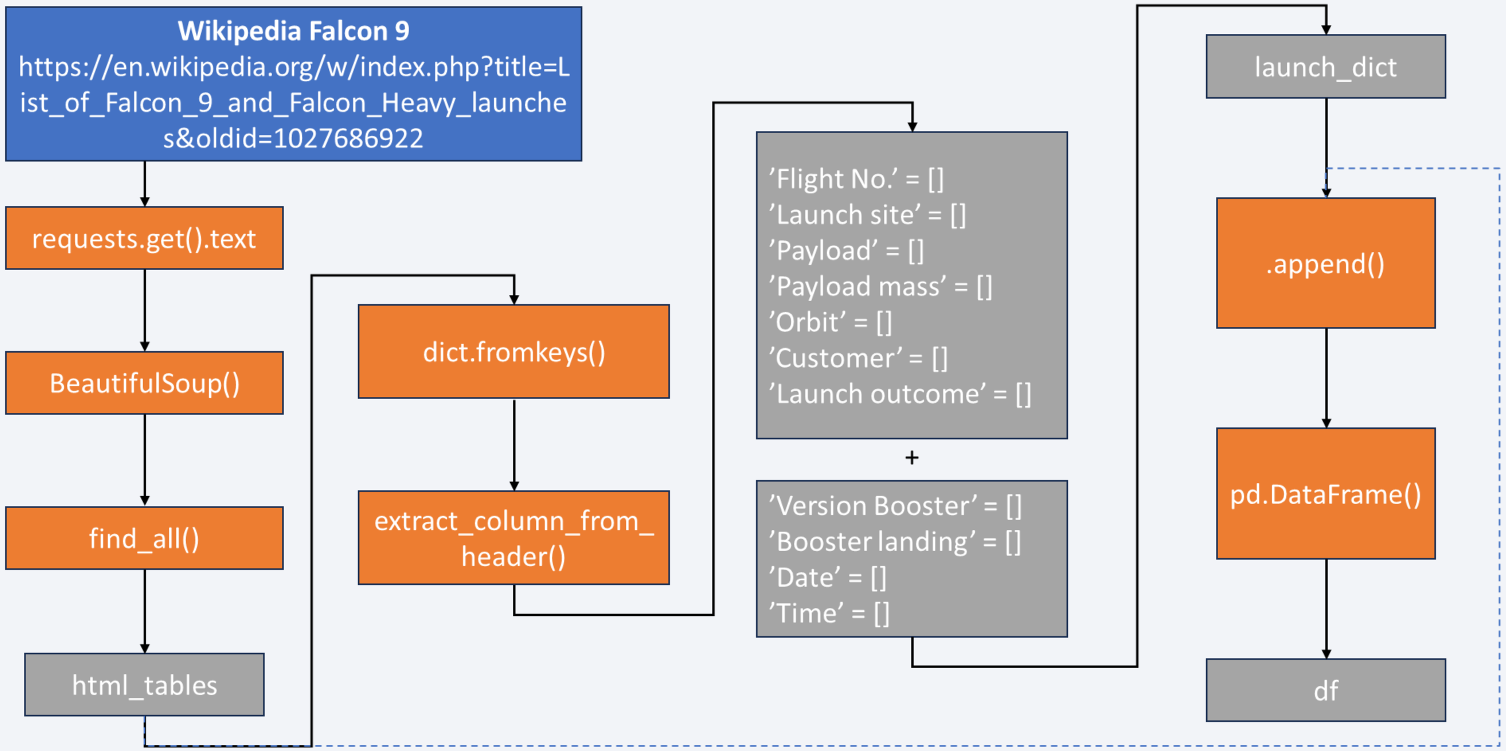
Data Collection – SpaceX API

- A request was made to the SpaceX API using the **requests** library
- The received .json was flattened into a **Pandas** data frame using pd.json_normalize()
- The data was subset to include only data on some parameters
- This data was sourced using IDs for each launch and four user-defined functions (getBoosterVersion(), getLaunchSite(), getPayloadData(), getCoreData()) to construct the dataframe df
- df was filtered to reach final data frame containing data for Falcon 9 launches
- https://github.com/Rcubed19/spacex/blob/main/Lab1_collecting_data.ipynb



Data Collection - Scraping

- The contents for Falcon 9 launches was requested from Wikipedia (static page) as BeautifulSoup
- The BS was extracted for column and variable names
- The column and variable names were used to create a dictionary
- This dictionary was transformed into a Pandas data frame called df
- https://github.com/Rcubed19/spacex/blob/main/Lab2_webscraping.ipynb



Data Wrangling

- The data collected in the last phase was examined, missing values in each attribute were identified
- 'LandingPad' was identified to be data type 'object' with 28% missing values
- The number of launches for each launch site were calculated by using the method .value_counts() on the variable 'LaunchSite'
- Next the number and occurrence of each target orbit were determined similarly with the method '.value_counts()' on the variable 'Orbit'
- Mission outcomes per target orbit type were determined with the method '.value_counts()' on the variable 'Outcome'. The outcome 'True Ocean' means successful landing to a specific region of the sea, while 'False Ocean' means unsuccessful landing to sea. 'True RTLS' means successful return to launch site and a successful ground pad landing; 'False RTLS' is an unsuccessful event. 'True ASDS' is a successful mission outcome with landing to a drone ship, 'False ASDS' means unsuccessful landing to a drone ship. 'None ASDS' and 'None None' are failed landings
- Unsuccessful landing types were collected into a new variable 'bad_outcomes'

Data Wrangling (continued)

- Next, by using the ‘Outcome’ variable, a new variable ‘landing_class’ was created with bad outcomes with value 0 and successful outcomes with value 1
- Now, using ‘landing_class’ we can create a new variable to the data frame ‘Class’ with 0 for unsuccessful outcomes and 1 for successful ones by simply: df[‘Class’] = landing_class
- Finally, using ‘Class’ we can calculate the overall success rate for all SpaceX Falcon 9 launches with df[‘Class’].mean()
- We find the overall success rate to be 66.6%
- https://github.com/Rcubed19/spacex/blob/main/Lab3_data_wrangling.ipynb

EDA with Data Visualization

- Exploratory data analysis (EDA) was performed using Pandas and Matplotlib
- The following graphs were made:
  - flight number vs. launch site
  - payload vs launch site
  - success rate vs orbit type
  - flight number vs orbit type
  - payload vs orbit type
  - success rate yearly trend
- https://github.com/Rcubed19/spacex/blob/main/Lab4_EDA_and_data_visualization.ipynb

EDA with SQL

- After loading and establishing a connection with the database, a few queries were made with SQL (see table)

- https://github.com/Rcubed19/spacex/blob/main/Lab5_EDA_with_SQL.ipynb

Query	Output
select distinct Launch_Site from SPACEXTABLE	Unique launch sites
select sum(PAYLOAD_MASS__KG_) as total_payload_mass from SPACEXTABLE where Customer = 'NASA (CRS)'	Total payload mass for NASA CRS missions
select avg(PAYLOAD_MASS__KG_) as average_payload_mass from SPACEXTABLE where Booster_Version = 'F9 v1.1'	Average payload mass carried by Falcon 9 (v. 1.1)
select min(Date) from SPACEXTABLE where Landing_Outcome = 'Success (drone ship)'	First successful landing on a drone ship
select Booster_Version, Payload, PAYLOAD_MASS__KG_ from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000	The names of the boosters which had a successful landing to ground pads with payload mass between 4000-6000 kg
select Mission_Outcome, count(Mission_Outcome) as Outcome from SPACEXTABLE group by Mission_Outcome	Total number of successful and failed mission outcomes
select Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)	Booster versions with maximum payload mass
select Date from SPACEXTABLE where landing_outcome = 'Success (ground pad)' and substr(Date,0,5),='2017'	List the successful landing outcomes to ground pads in months in 2017
select Landing_Outcome, count(*) as Count_Cases from SPACEXTABLE where Date between '2010-06-04' and '2017-03-20' group by Landing_Outcome order by Count_Cases DESC	Rank the count of landing outcomes between 2010-06-04 and 2017-03-20 in descending order

Build an Interactive Map with Folium

- An interactive map was constructed using Folium
- This map marked the different launch sites used by SpaceX together with the number of successful launches and distance from different close-by features
- This is because the launch success rate may depend on the location and proximities of a launch site, i.e., the initial position of rocket trajectories.
- https://github.com/Rcubed19/spacex/blob/main/Lab6_folium.ipynb

Build a Dashboard with Plotly Dash

- An interactive dashboard was built using Plotly Dash with two different plots
- A pie chart of successful launches with a selection tool for single launch sites (success vs. failure) vs. total successful launches per launch site
  - This shows nicely how successful launces are distributed across launch sites
- A scatter plot of payload mass vs. launch success with sliding tool to determine payload mass range. Booster version as additional info.
  - This shows how success is distributed across different payload masses and booster versions
- https://github.com/Rcubed19/spacex/blob/main/Lab7_spacex_dash_app.py

Predictive Analysis (Classification)

- Exploratory data analysis was performed using classification with the sklearn library.
- The following steps were done during the work
  - A column for the outcome (Class) was created by applying the method to_numpy()
  - The data was standardized with preprocessing.StandardScaler() and split into training and test data using train_test_split() (test size 0.2, random state = 2)
  - Hyperparameters were optimized for support vector machine (SVM), classification tree, K-nearest neighbor and logistic regression using the training data
  - The best method for classification was found using test data
- https://github.com/Rcubed19/spacex/blob/main/Lab8_landing_prediction.ipynb

Results

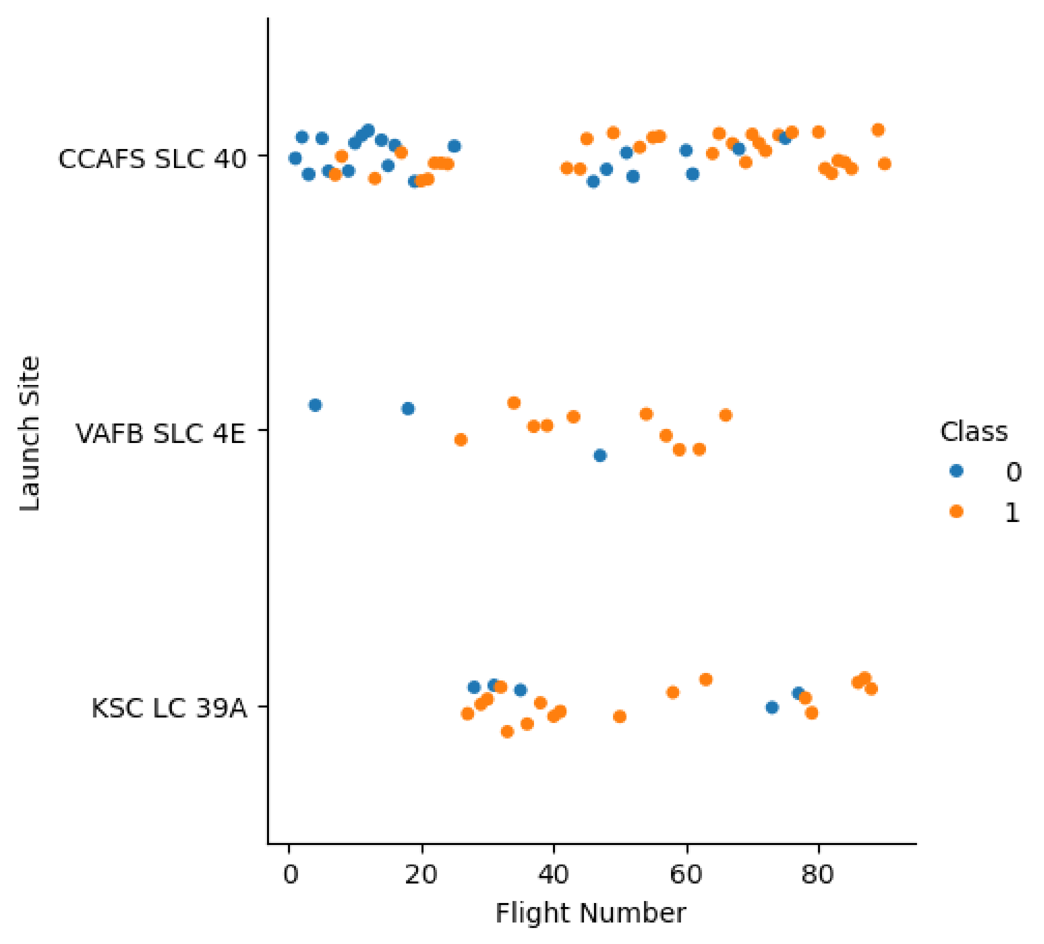
- In the next slides we'll present
  - Results from our exploratory data analysis using Pandas, Matplotlib and SQL
  - Screenshots from interactive analytics using Folium and Plotly dash
  - Predictive analysis results for four different ML models (logistic regression, KNN, decision tree and SVM) using sklearn

Section 2

Insights drawn from EDA

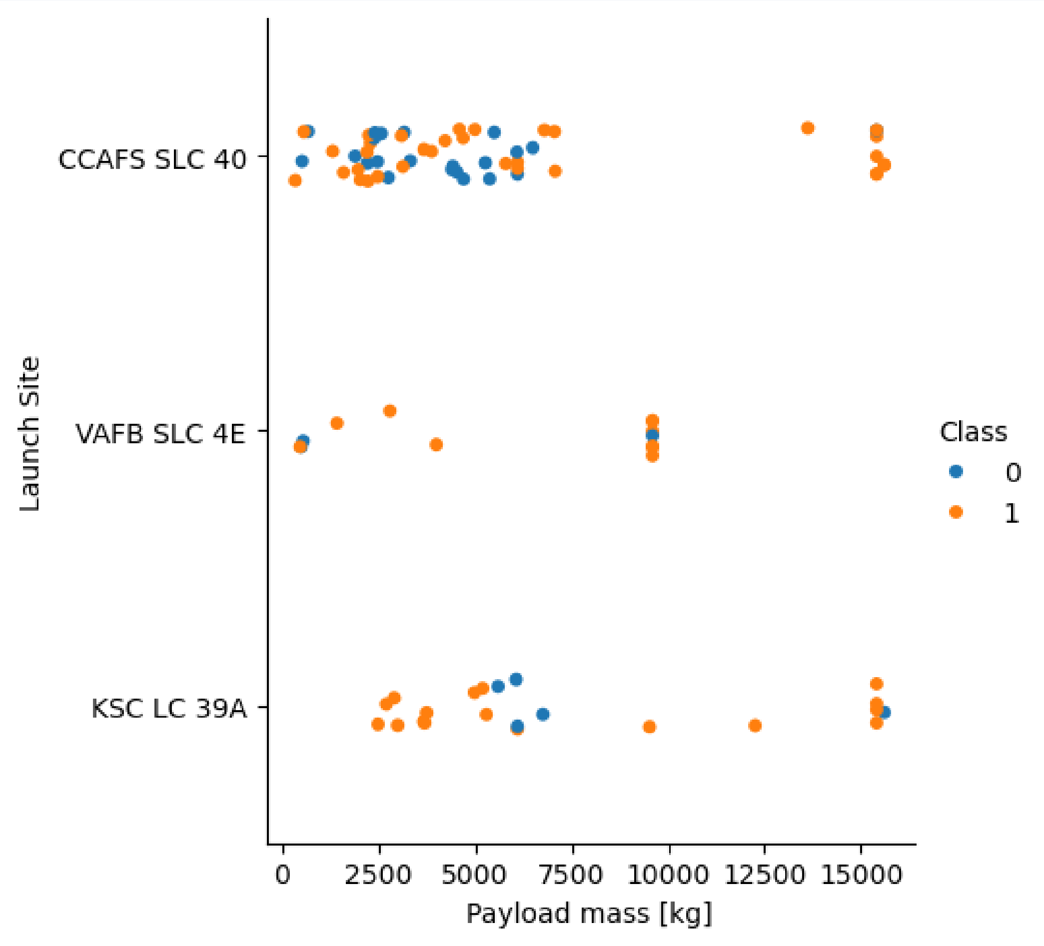
Flight Number vs. Launch Site

- We can see how operations at different launch sites have succeeded over time (increasing flight number)
- CCAFS SLC 40 has most launches
- KSC LC39A and VAFB SLC 4E have higher ratio of successful launches (class = 1)



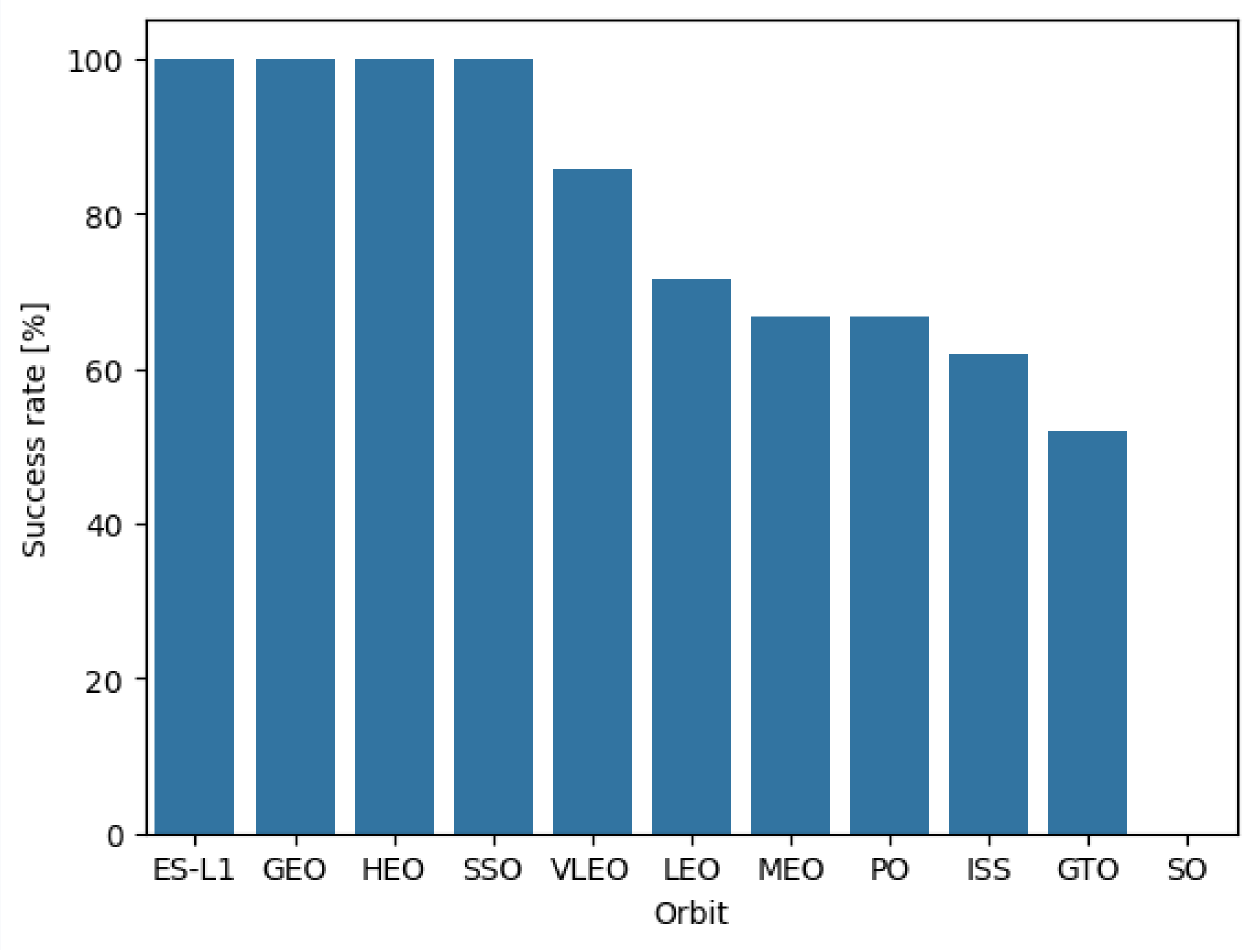
Payload vs. Launch Site

- Most small payload launches have occurred at CCAFS SLC 40 which also present most unsuccessful launches



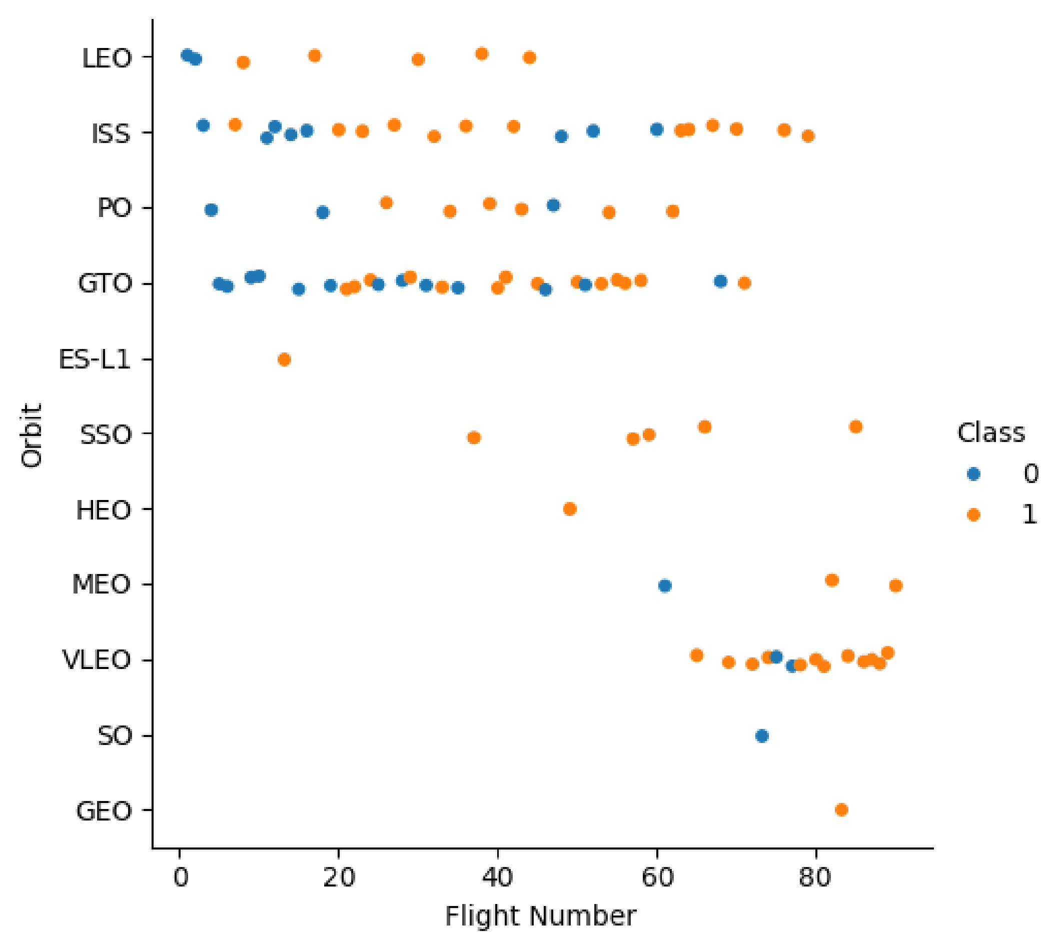
Success Rate vs. Orbit Type

- There's variation in the success rates for different orbits:
  - 100% of launches to ES-L1, GEO, HEO, SSO have been successful
  - 0% of launches to SO have been successful
- These extreme cases represent only a minority of all launches



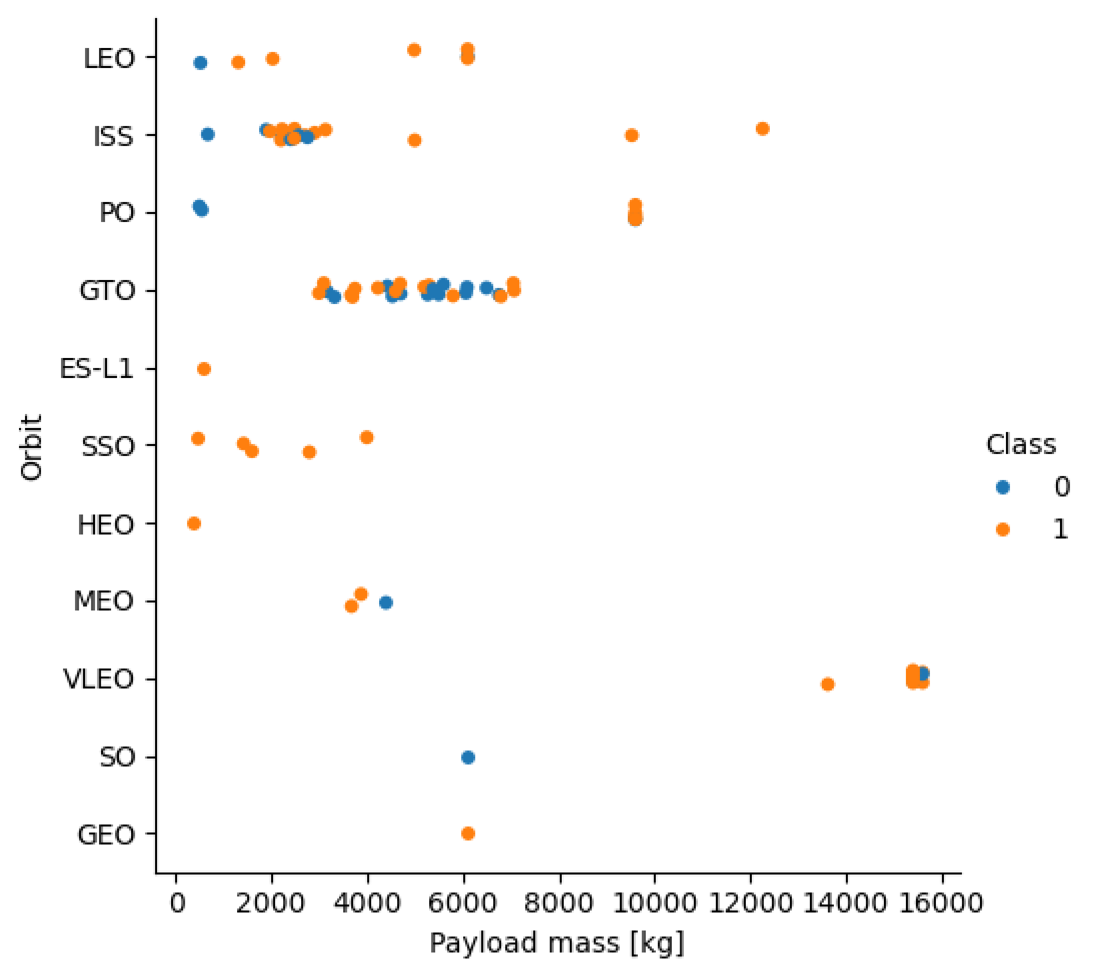
Flight Number vs. Orbit Type

- The initial launches (low flight number) were to LEO, ISS, PO and GTO
- Later, most launches were to VLEO (flight number > 60)
- Most unsuccessful flights were to GTO, ISS



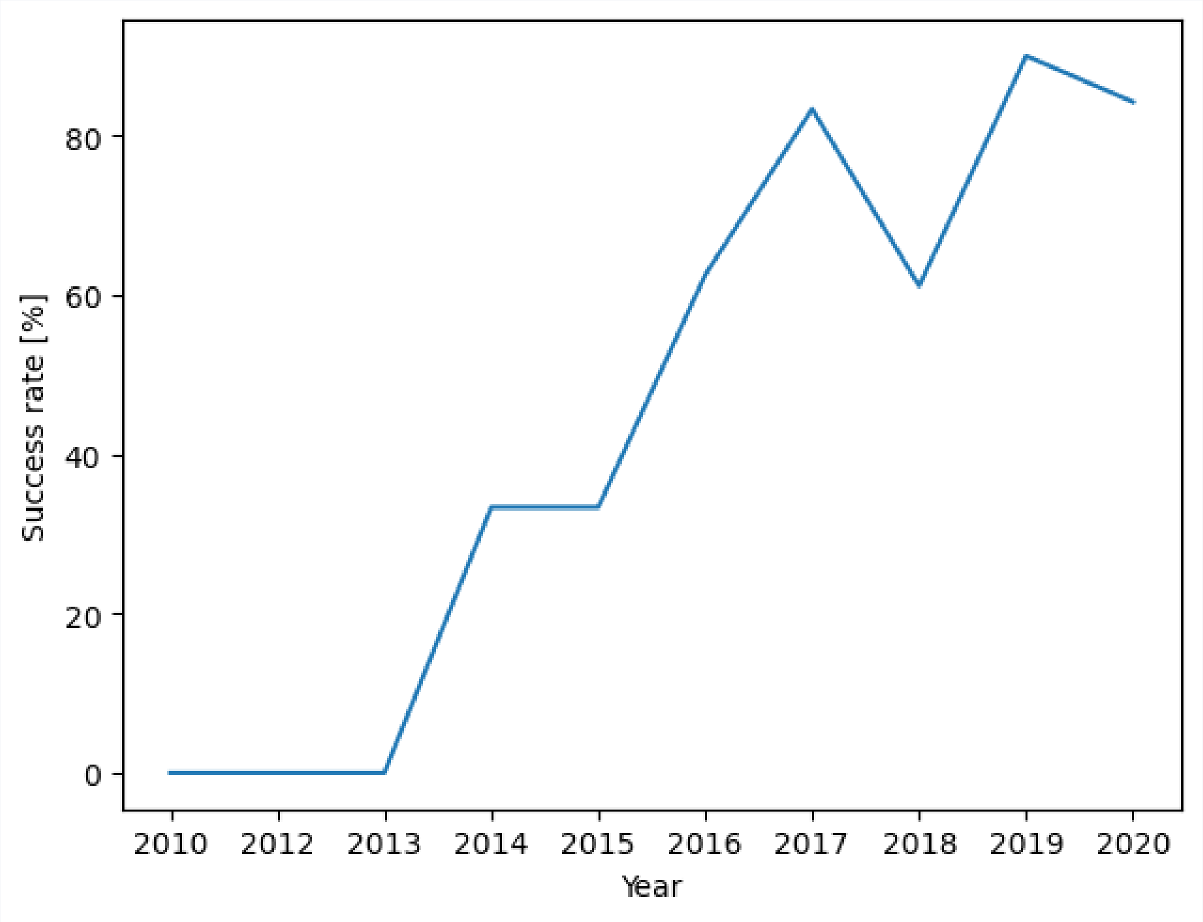
Payload vs. Orbit Type

- The widest range of payload masses have been launched to ISS
- All GTO launches have been between 2000 and 8000 kg
- All VLEO launches have been >12000 kg while LEO launches have been ≤ 6000 kg



Launch Success Yearly Trend

- The average success rate has been steadily increasing from 2013 until 2017
- 2019 onwards the success rate has been ≥ 80%



All Launch Site Names

- The names of unique launch sites can be found using a query containing:
  - DISTINCT operator together with SELECT

```
%sql select distinct Launch_Site from SPACEXTABLE
```

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

Launch Site Names Begin with 'KSC'

- Launch sites (first 5) can be queried using a query containing:
  - SELECT * (select all) from SPACEXTABLE
  - Use the WHERE and LIKE operators to limit to cases where ‘Launch_Site’ begins with ‘KSC’

```
%sql select * from SPACEXTABLE where Launch_Site like 'KSC%' limit 5
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-03-16	06:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-01-05	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
2017-05-15	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

Total Payload Mass

- The total payload mass carried by boosters for NASA CRS can be calculated with a query containing the following:
  - SELECT PAYLOAD_MASS_KG from SPACEXTABLE
  - Select only cases where customer is ‘NASA (CRS)’ with the WHERE operator
  - Give this the name ‘total_payload_mass’ with the AS operator
- Like this:

```
%sql select sum(PAYLOAD_MASS__KG_) as total_payload_mass from SPACEXTABLE where Customer = 'NASA (CRS)'
```

total_payload_mass
45596

Average Payload Mass by F9 v1.1

- To calculate the average payload mass carried by Falcon 9 booster version 1.1 we use:
  - AVG() to calculate the average ‘PAYLOAD_MASS_KG’
  - AS operator to give the output the name ‘average_payload_mass’
  - WHERE operator to only select cases where “Booster_Version == ‘F9 v.1.1’”

```
%sql select avg(PAYLOAD_MASS__KG_) as average_payload_mass from SPACEXTABLE where Booster_Version = 'F9 v1.1'
```

average_payload_mass
2928.4

First Successful Drone Ship Landing Date

- The date for the first successful drone ship landing can be found with a query containing:
  - WHERE operator to select the cases where “Landing_Outcome == ‘Success (drone ship)’”
  - MIN() operator to select the earliest case

```
%sql select min(Date) from SPACEXTABLE where Landing_Outcome = 'Success (drone ship)'
```

min(Date)
2016-05-27

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

```
%sql select Booster_Version, Payload, PAYLOAD_MASS__KG_ from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000
```

Booster_Version	Payload	PAYLOAD_MASS__KG_
F9 FT B1032.1	NROL-76	5300
F9 B4 B1040.1	Boeing X-37B OTV-5	4990
F9 B4 B1043.1	Zuma	5000

Total Number of Successful and Failure Mission Outcomes

- We can calculate the total number of successful and unsuccessful mission outcomes with a query containing:
  - COUNT and AS operators to tally up the type of 'Mission_Outcome' as 'Outcome'
  - GROUP BY operator to group the output by 'Mission_Outcome'

```
%sql select Mission_Outcome, count(Mission_Outcome) as Outcome from SPACEXTABLE group by Mission_Outcome
```

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

Boosters Carried Maximum Payload

- The names of the Boosters that have carried the maximum payload mass can be found with a query containing the following:
  - WHERE and MAX() operators to select the largest PAYLOAD_MASS_KG
  - SELECT to select the Booster_Version

```
%sql select Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```

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

2017 Launch Records

- The successful landings on ground pad in 2017 can be found with the following query:
  - WHERE operator together with SUBSTR() to select the cases where year in ‘Date’ is 2017 and “Landing_Outcome == ‘Success (ground pad)’”
  - SUBSTR() operator to select month from ‘Date’
  - SELECT and AS operators to select month, Booster_Version, Launch_Site, Payload, PAYLOAD_MASS_KG, Mission_Outcome, Landing_Outcome

```
%sql select substr(Date,0,5) as year, substr(Date, 6, 2) as month, Booster_Version, Launch_Site, Payload, PAYLOAD_MASS__KG_, Mission_Outcome, Landing_Outcome from SPACEXTABLE where substr(Date,0,5)='2017' and Landing_Outcome = 'Success (ground pad)'
```

year	month	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Mission_Outcome	Landing_Outcome
2017	02	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	Success	Success (ground pad)
2017	01	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	Success	Success (ground pad)
2017	03	F9 FT B1035.1	KSC LC-39A	SpaceX CRS-11	2708	Success	Success (ground pad)
2017	08	F9 B4 B1039.1	KSC LC-39A	SpaceX CRS-12	3310	Success	Success (ground pad)
2017	07	F9 B4 B1040.1	KSC LC-39A	Boeing X-37B OTV-5	4990	Success	Success (ground pad)
2017	12	F9 FT B1035.2	CCAFS SLC-40	SpaceX CRS-13	2205	Success	Success (ground pad)

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

- We can find the landing outcomes between two dates (2010-06-04 and 2017-03-20) with a query containing:
  - COUNT() operator to count the cases as 'Count_cases'
  - WHERE , BETWEEN and AND operators to select 'Date' between two values
  - GROUP BY and ORDER BY with DESC to order the 'Landing_Outcome' by the 'Count_cases' in descending order

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

```
%sql select Landing_Outcome, count(*) as Count_Cases from SPACEXTABLE where Date between '2010-06-04' and '2017-03-20' group by Landing_Outcome order by Count_Cases DESC
```

Section 3

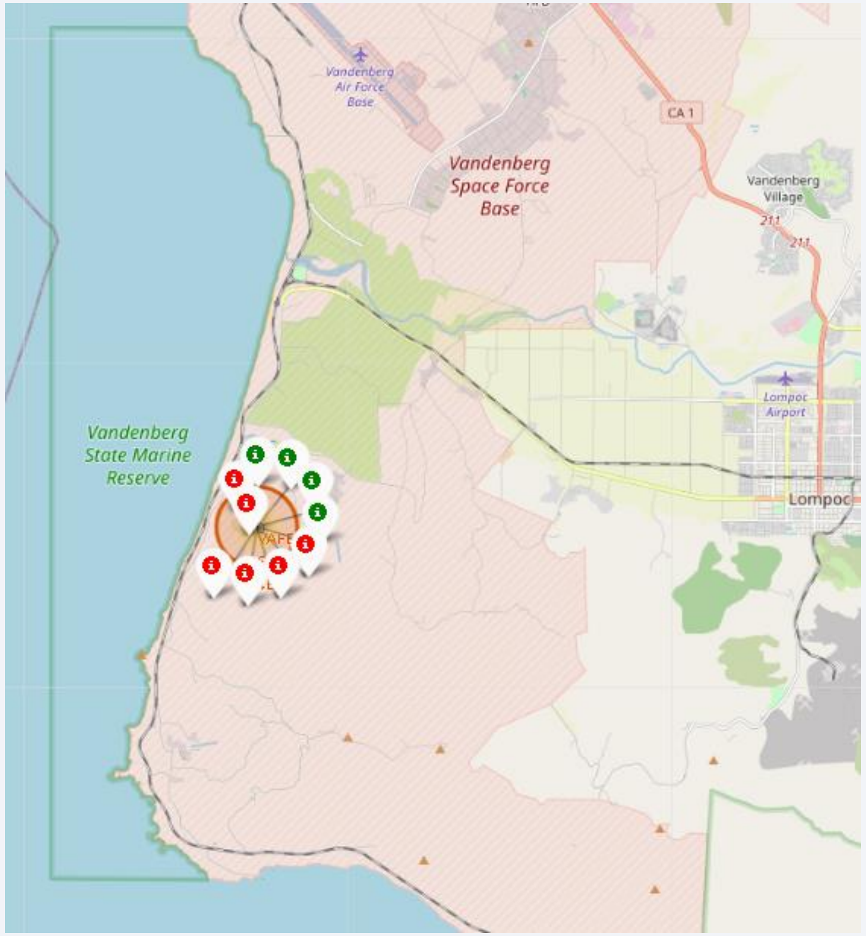
Launch Sites Proximities Analysis

Folium: SpaceX launches across the US

- From our Folium implementation, it is clear how SpaceX conducts launches from both West and East coast of the US
- Most launches are from Florida (total of 46) where the Kennedy Space Center (KSC LC-39A) and Cape Canaveral Space Force Station (CCAFS SLC-40 & CCAFS LC-40)
- On the west coast, only Vandenberg Space Force Station (VAFB LC-4E) sustains SpaceX Falcon 9 launches

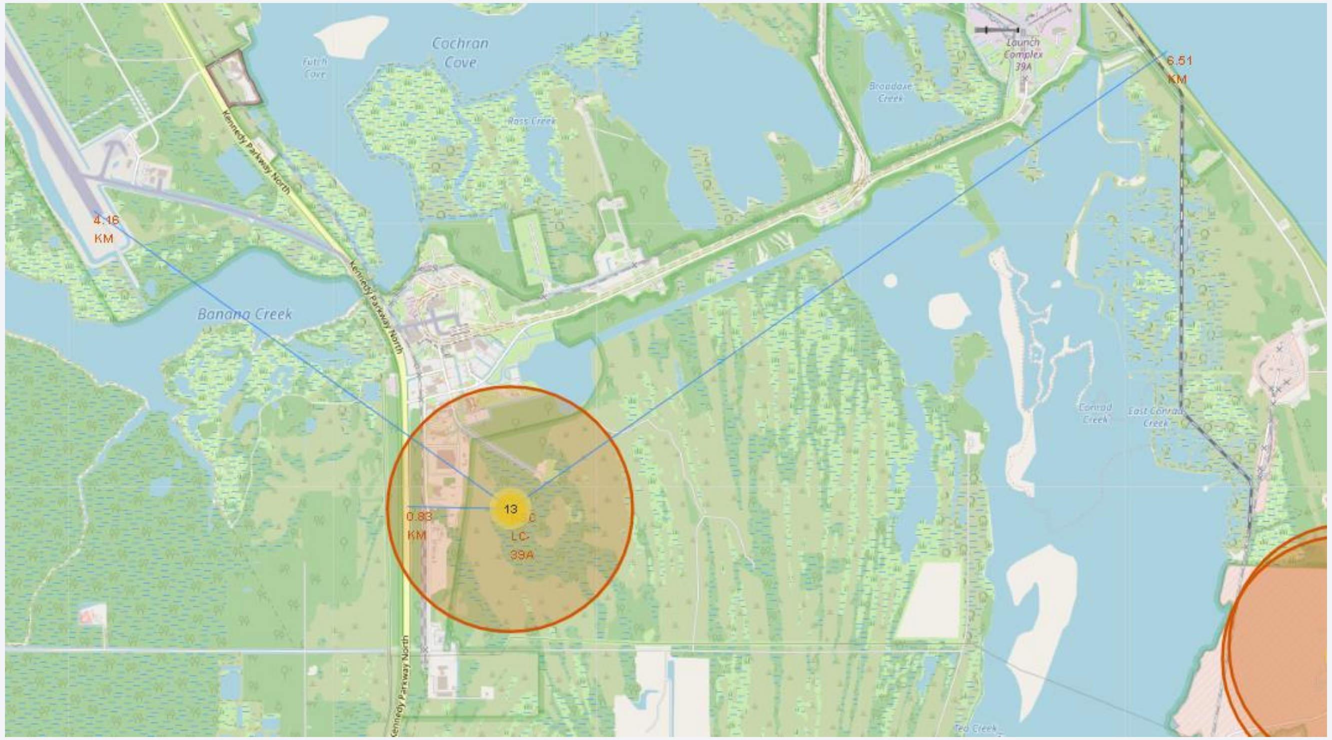
Folium: Success/failure at VAFB SLC-4E

- Our Folium implementation allows us to look at the data at launch site-level
- For example, here we have a closer look at the successful/failed launches from Vandenberg (VAFB SLC-4E) shown in green/red



Folium: Proximities from LC-39A

- We can explore the proximities for different launch sites, e.g. LC-39A
- We can find that the
    - Distance to seashore is just over 6.5 km
    - The highway is only 0.83 km away
    - The nearest runway is 4.16 km away
- The runway and highway allow for transport to the launch site
- The launch trajectory will take the rockets towards the sea allowing for safe launches for the community

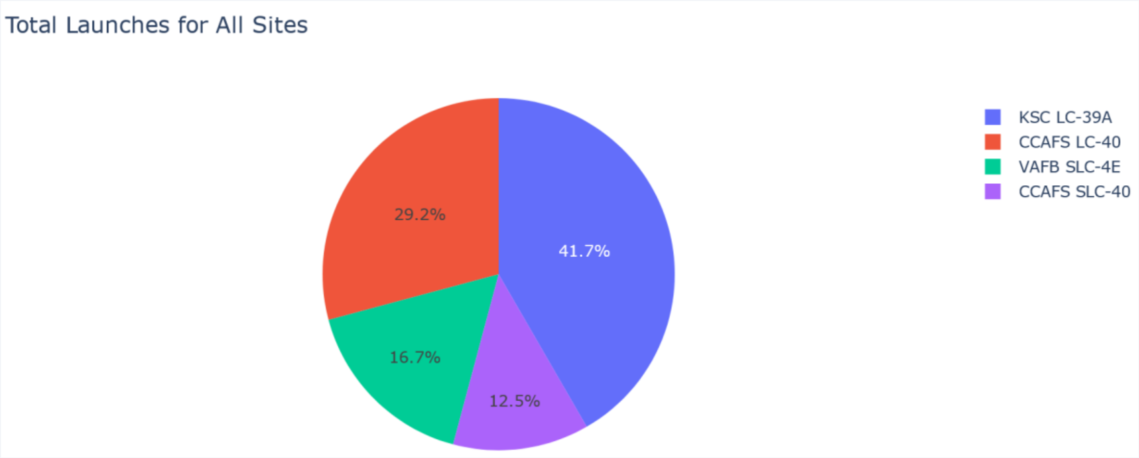


Section 4

Build a Dashboard with Plotly Dash

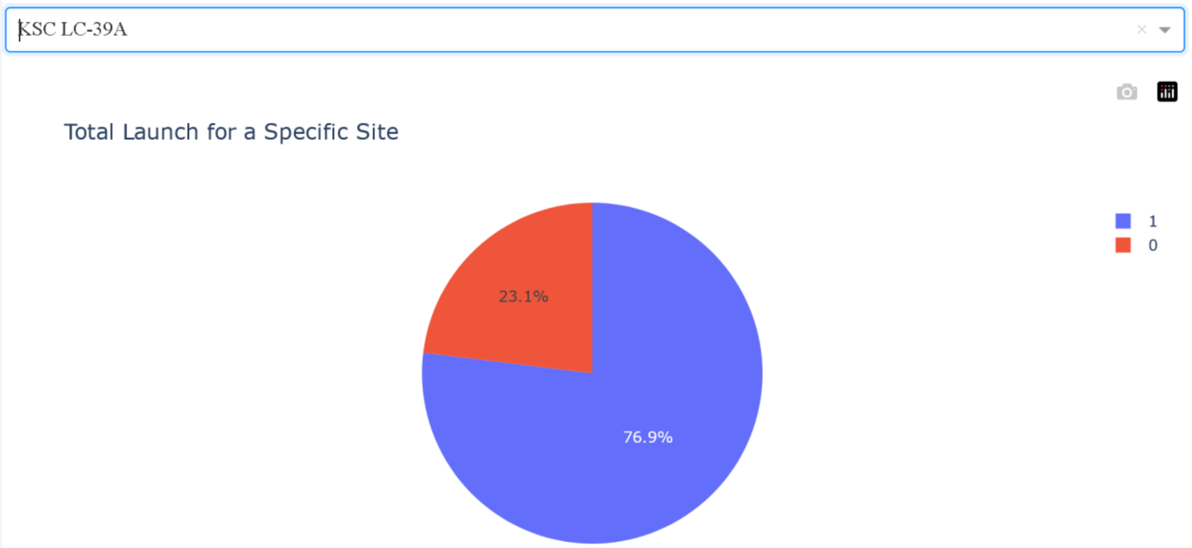
Plotly dash: total successful launches

- On our interactive Plotly dash we can visualize how the the successful launches are distributed across the different launch sites
- Of the total successful launches 41.7% were from KSC LC-39A



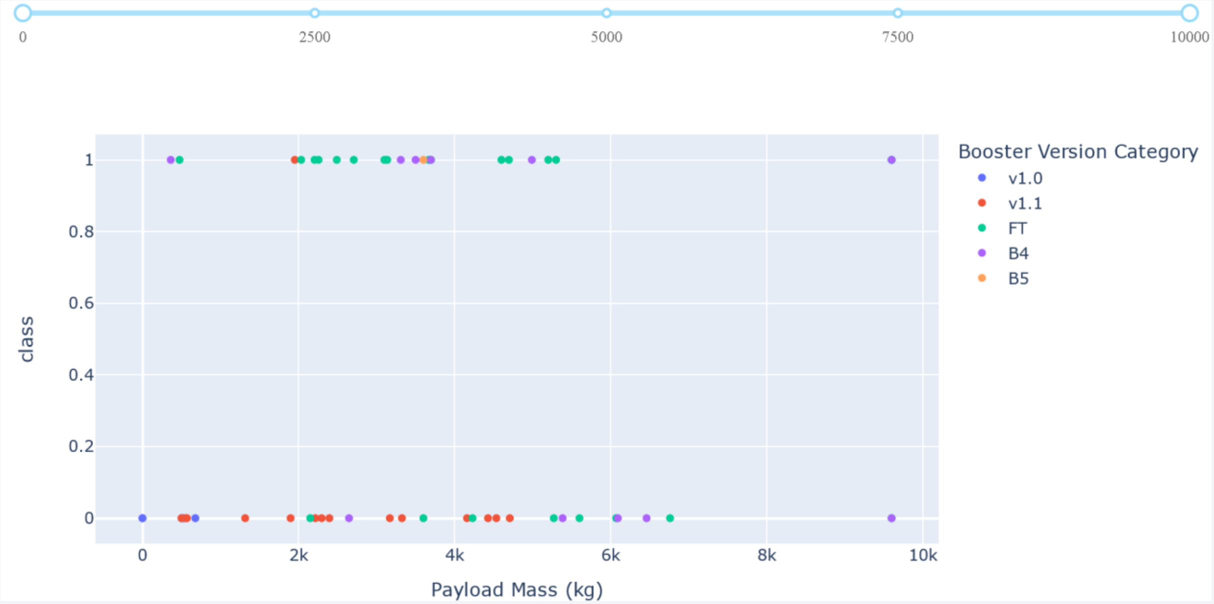
Plotly dash: Ratio of successful and failed launches

- We look at the ratio of successful and failed launches by launch site with a pie chart
- The highest ratio of successful to failed launches were at KSC LC-39A
- Successful launches are show as blue and failed as red



Plotly dash: Launch outcome by payload

- Another interesting look at the data is by plotting launch outcome (0 to 1 on y-axis) to payload mass (x-axis) as a scatter plot
- The different colors show different booster versions
- On top we have a slider allowing us to select a range of payload masses shown on the scatter plot
- Here outcome class 0 contains failed launches and it appears that this has more spread on the payload mass than the successful ones (class 1)



Section 5

Predictive Analysis (Classification)

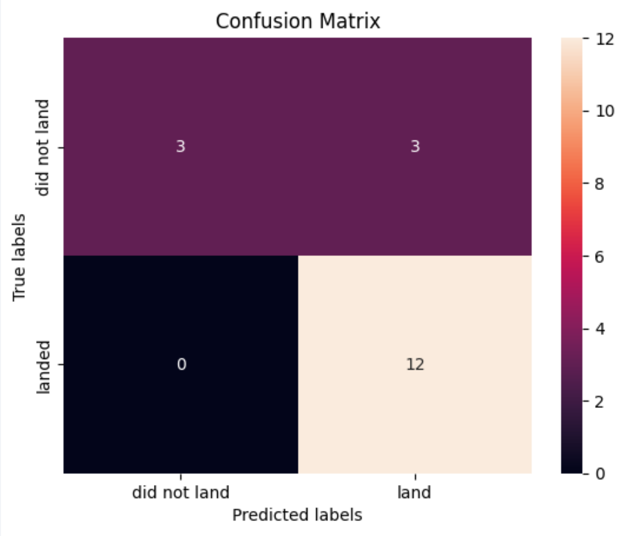
Classification Accuracy

- Three of the four tested models (SVM, KNN and logistic regression) all had the same classification accuracy
- Classification tree performed the poorest

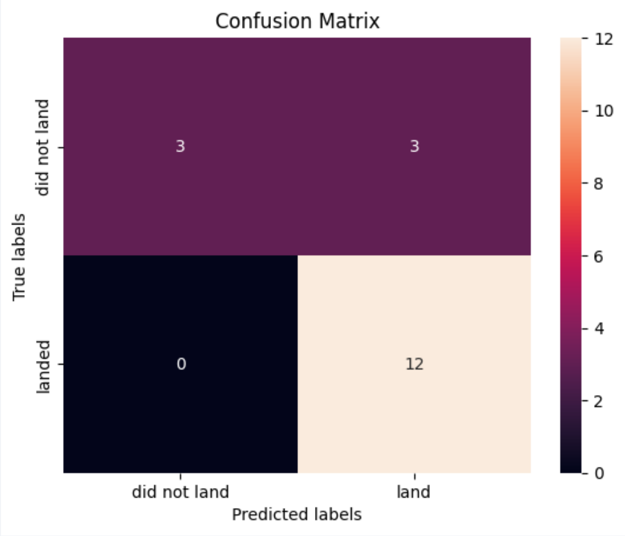
Confusion Matrix

- For all the best performing models, the confusion matrix is the same: all true landed cases were predicted correctly.
- However, for the 6 cases that did not, half were predicted wrongly to have landed correctly

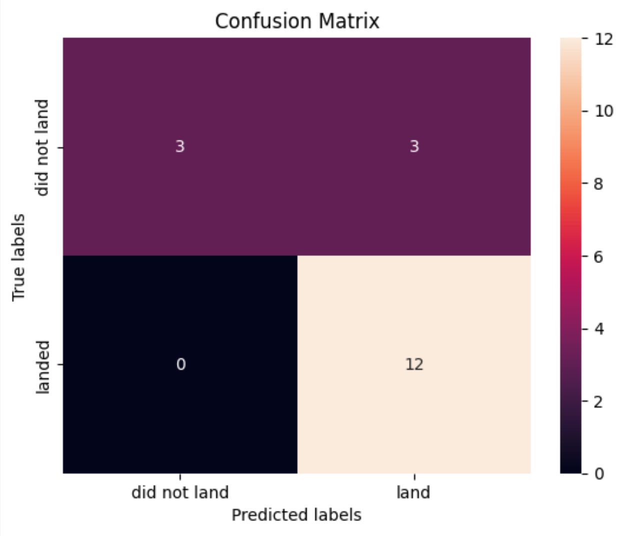
Logistic regression



SVM



KNN



Conclusions

- Concluding, we find that
  - Success rate has improved over the years (increasing flight number)
  - The results of the (EDA) exploratory data analysis revealed the success rate of the SpaceX Falcon 9 rocket landings is 66%.
  - Some launches on specific orbits (ES-L1, GEO, HEO, and SSO) have a 100% success rates but these are single events
  - Successful launch locations are found both at west coast and east coast locations in the US
  - All tested ML algorithms (logistic regression, KNN, SVM) except for decision tree yielded 83% accuracy on test data

Thank you!