IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jiayi Pan
January 1, 2024

Outline

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

Executive Summary

In this project, SpaceX launch data are collected through its website using API request and web scraping methods. The data are wrangled by replacing miss-values by the mean, and a column showing successes of booster recovery is generated. SQL is used to query various info in the data table. Launch site locations are visualized and analyzed with Folium map, and a dashboard application with Plotly Dash is developed to display the launch characteristics interactively. The Exploratory Data Analysis reveals important relationship between SpaceX rocket launch parameters. The machine learning models are constructed to predict the outcomes of the launches. These analyses suggest 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%. Since 2013, the success rate kept increasing till 2020. The Decision Tree has the highest accuracy (R^2) of 0.89 in predicting the outcome of a rocket launch.

Introduction

The commercial space age is here, companies are making space travel affordable for everyone. Virgin Galactic is providing suborbital spaceflights. Rocket Lab is a small satellite provider. Blue Origin manufactures sub-orbital and orbital reusable rockets. Perhaps the most successful is SpaceX. SpaceX's accomplishments include: sending spacecraft to the International Space Station; building up starlink, a satellite internet constellation providing satellite Internet access; and sending manned missions to Space. One reason SpaceX can do this is the rocket launches are relatively inexpensive. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. Sometimes the first stage does not land. Sometimes it will crash as shown in this clip. Other times, Space X will sacrifice the first stage due to the mission parameters like payload, orbit, and customer. In this capstone, we will try to determine the price of each launch, by gathering information about Space X and creating dashboards. We will also determine if SpaceX will reuse the first stage. Instead of using rocket science to determine if the first stage will land successfully, we will train machine learning models and use public information to predict if SpaceX will reuse the first stage.

Section 1

Methodology

Methodology

Executive Summary

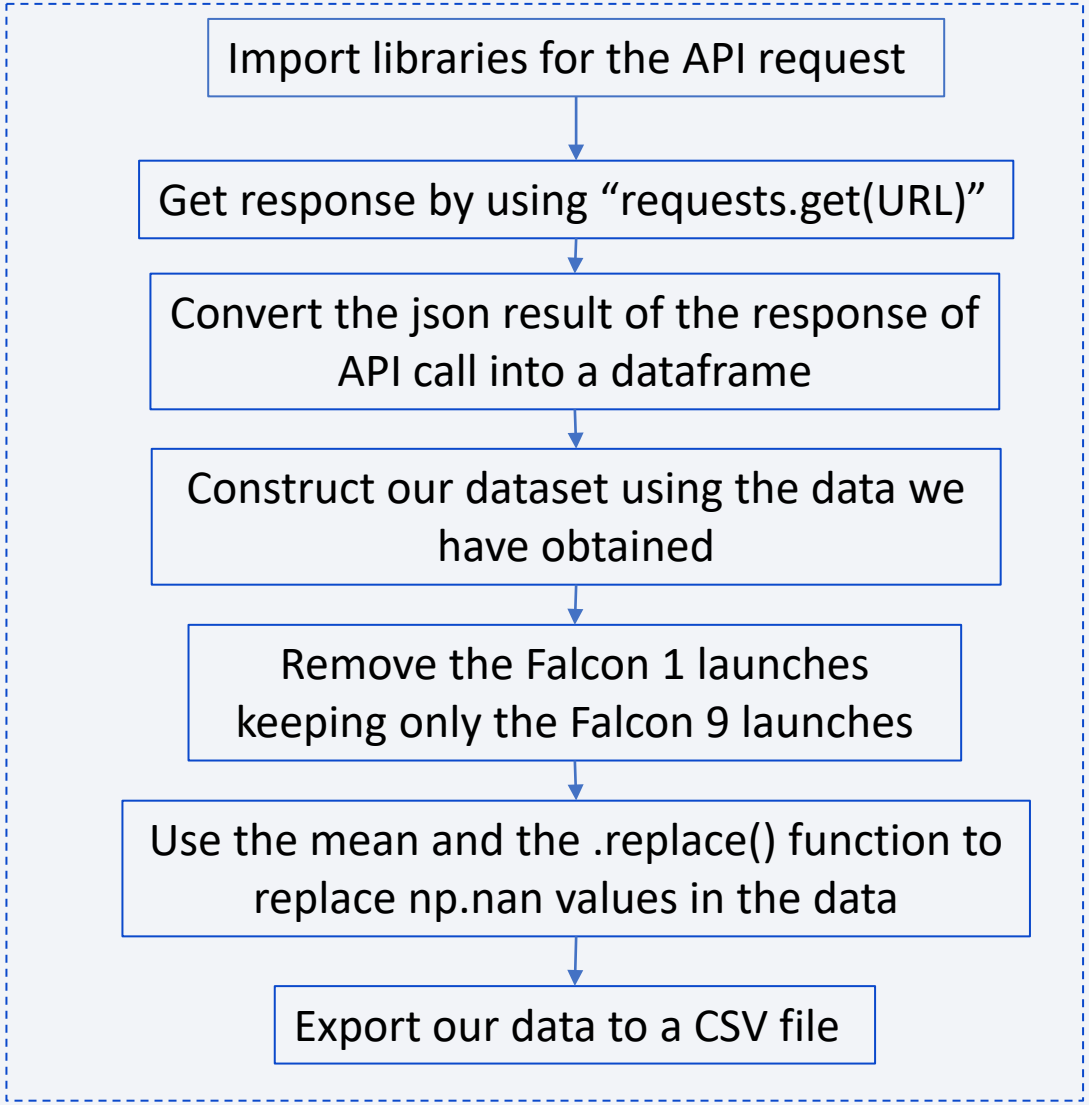
- Data collection:
  - The data are collected through API requests on the SpaceX website.
- Perform data wrangling
  - The missing data of payload mass are filled average value.
  - A landing outcome label is created showing the landing status.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The independent data are standardized. The four machine learning models are used to predict the success rate. The optimum parameters are sought through GridSearchCV method.

Data Collection

- Data are collected from SpaceX website through API requests.
- The detailed processes are described as follows.

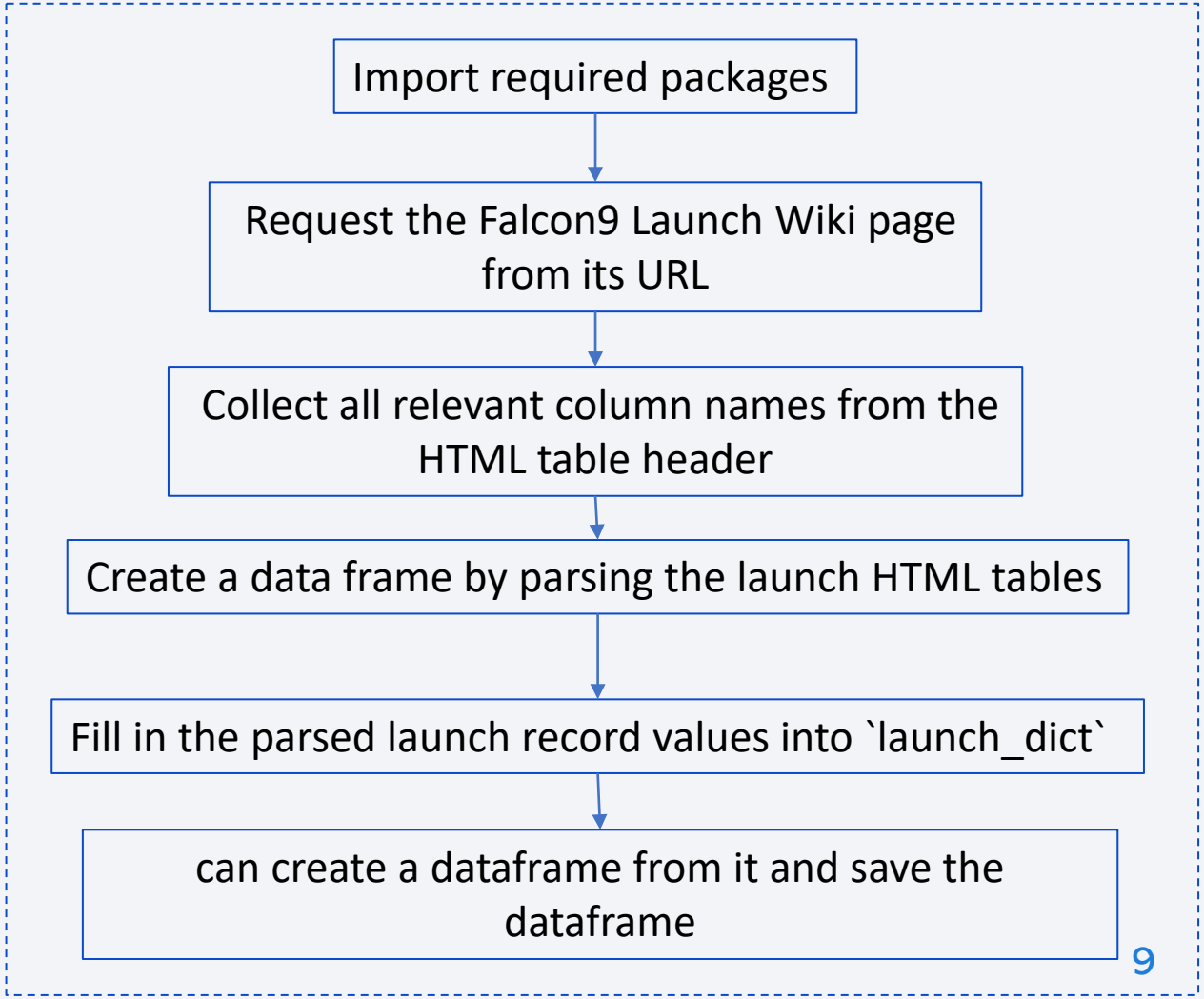
Data Collection – SpaceX API

- Request and parse the SpaceX launch data using the API GET request and create a datafame.
- Filter the dataframe to only include `Falcon 9` launches
- Dealing with Missing Values
- Save the data in a local repository
- https://github.com/panj1963/Applied_Data_Science_Capstone/blob/main/Lab_1.1.ipynb



Data Collection - Scraping

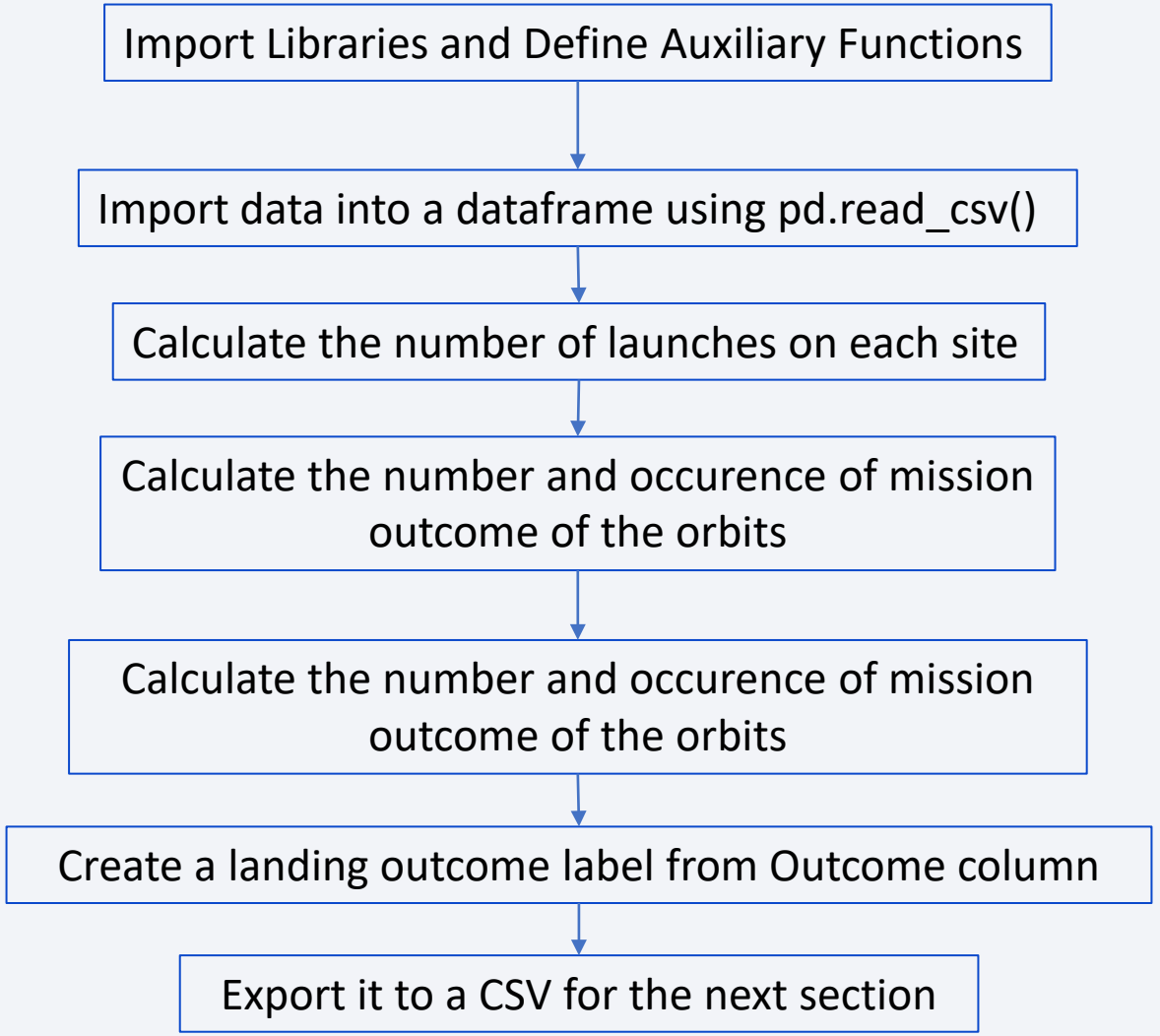
- Request the Falcon9 Launch Wiki page from its URL
- Extract all column/variable names from the HTML table header
- Create a data frame by parsing the launch HTML tables
- Save the data in a local repository
- https://github.com/panj1963/Applied_Data_Science_Capstone/blob/main/Lab_1.2.ipynb



Data Wrangling

- Calculate the number of launches on each site
- Calculate the number and occurrence of each orbit
- Calculate the number and occurence of mission outcome of the orbits
- Create a landing outcome label from Outcome column
- Export it to a CSV for the next section

- https://github.com/panj1963/Applied_Data_Science_Capstone/blob/main/Lab_1.3.ipynb



EDA with Data Visualization

- Generate a scatter plot of FlightNumber vs LaunchSite, showing the relationship between Flight Number and Launch Site.
- Generate a scatter plot of Launch sites vs their payload mass, showing relationship between Payload and Launch Site.
- Generate a barplot of Class vs Orbit type , showing the relationship between between success rate of each orbit type.
- Generate a scatter plot of FlightNumber vs Orbit type, showing relationship between FlightNumber and Orbit type.
- Generate a scatter plot of between Payload vs Orbit type, showing the relationship between Payload and Orbit type
- Generate a line plot of launch success rate vs time (year), showing yearly trend of launch success rate.
- https://github.com/panj1963/Applied_Data_Science_Capstone/blob/main/Lab_2.2.ipynb

EDA with SQL

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first succesful landing outcome in ground pad was acheived.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes

EDA with SQL

- List the names of the booster versions which have carried the maximum payload mass
- List the records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015.
- 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.

https://github.com/panj1963/Applied_Data_Science_Capstone/blob/main/Lab_2.1.ipynb

Build an Interactive Map with Folium

- Some circles, makers, and circle markers are created and added into a folium map
- Circles are used to show locations of NASA space centers. Makers are used to display the NASA space center names. Circle makers are used to display the launch successful rates.
- Add these objects to display info of the SpaceX launch info on a map.
- Add the GitHub URL of your completed interactive map with Folium map, as an external reference and peer-review purpose
- https://github.com/panj1963/Applied_Data_Science_Capstone/blob/main/Lab_3.1.ipynb

Build a Dashboard with Plotly Dash

- Generate a pie plots of the successful and non-successful rates at each launch site
- Generate scatter plots of class (successfulness) vs the payload mass
- Develop a dropdown to select sites: all sites or a specific site.
- Generate a RangeSlider to select a payload mass range.
- These plots and interactions are added to show these successful rates with selections of different sites and payload mass.
- https://github.com/panj1963/Applied_Data_Science_Capstone/blob/main/Lab_3.2.ipynb

Predictive Analysis (Classification)

Processes:

- Create a NumPy array from the column Class in data
- Standardize the data in X
- Split the data X and Y into training and test data.
- Create a logistic regression object then create a GridSearchCV object.
- Fit the object to find the best parameters from the dictionary parameters.
- Calculate the accuracy on the test data and draw the confusion matrix.
- Repeat the steps for developing model and implementing evaluation for support vector machine, decision tree, and k nearest neighbors methods.

Model refinement:

- Using the GridSearchCV methods to find the best parameters for a specific model, and using the scores and the confusion matrix to evaluate the model results.
- https://github.com/panj1963/Applied_Data_Science_Capstone/blob/main/Lab_4.1.ipynb

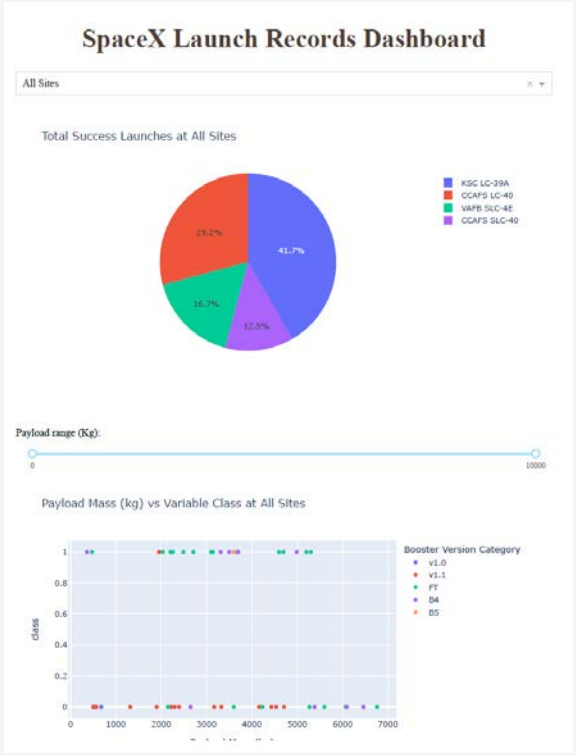
Results

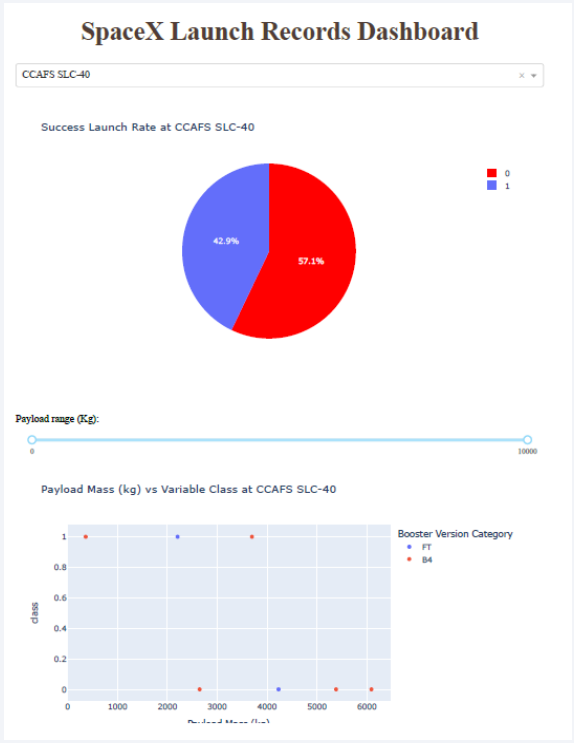
Exploratory data analysis results

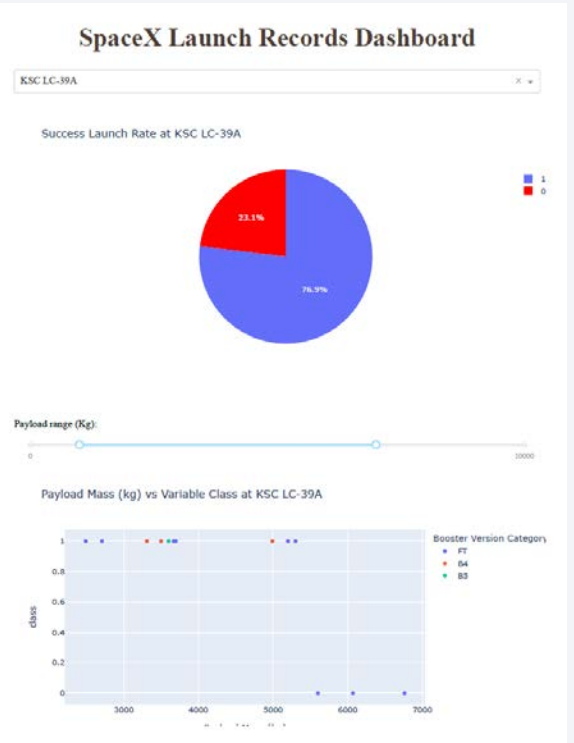
- Different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.
- For the VAFB-SLC launchsite there are no rockets launched for heavy payload mass (greater than 10000).
- For the LEO orbit, the success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.
- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS. However, for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.
- Since 2013, the success rate kept increasing till 2020.

Results

Interactive analytics demo in screenshots is displayed below. It illustrates how the successful rate is changed with launch sites and payload mass range.







Results

Predictive analysis results

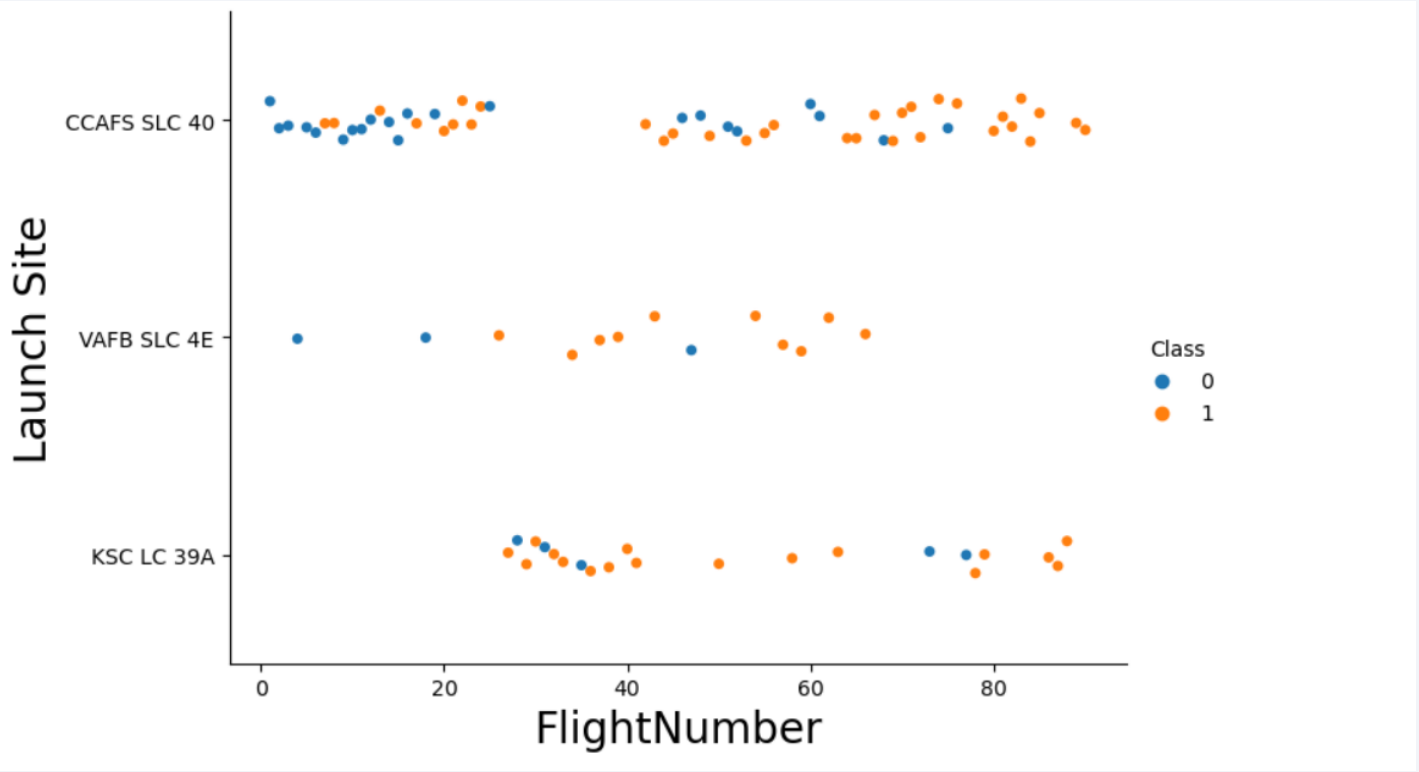
The Logistic Regression, Support Vector Machine, Decision Tree, and K Nearest Neighbors methods are used to predict the launch successful rate based on a number of parameters.

- For the four machine learning methods, the GridSearchCV can be used to find a set of best hyper-parameters.
- The Decision Tree has the highest accuracy (R^2) of 0.89 for the test dataset.
- The Logistic Regression, Support Vector Machine, and K Nearest Neighbors have same accuracy (R^2) of 0.83.

Section 2

Insights drawn from EDA

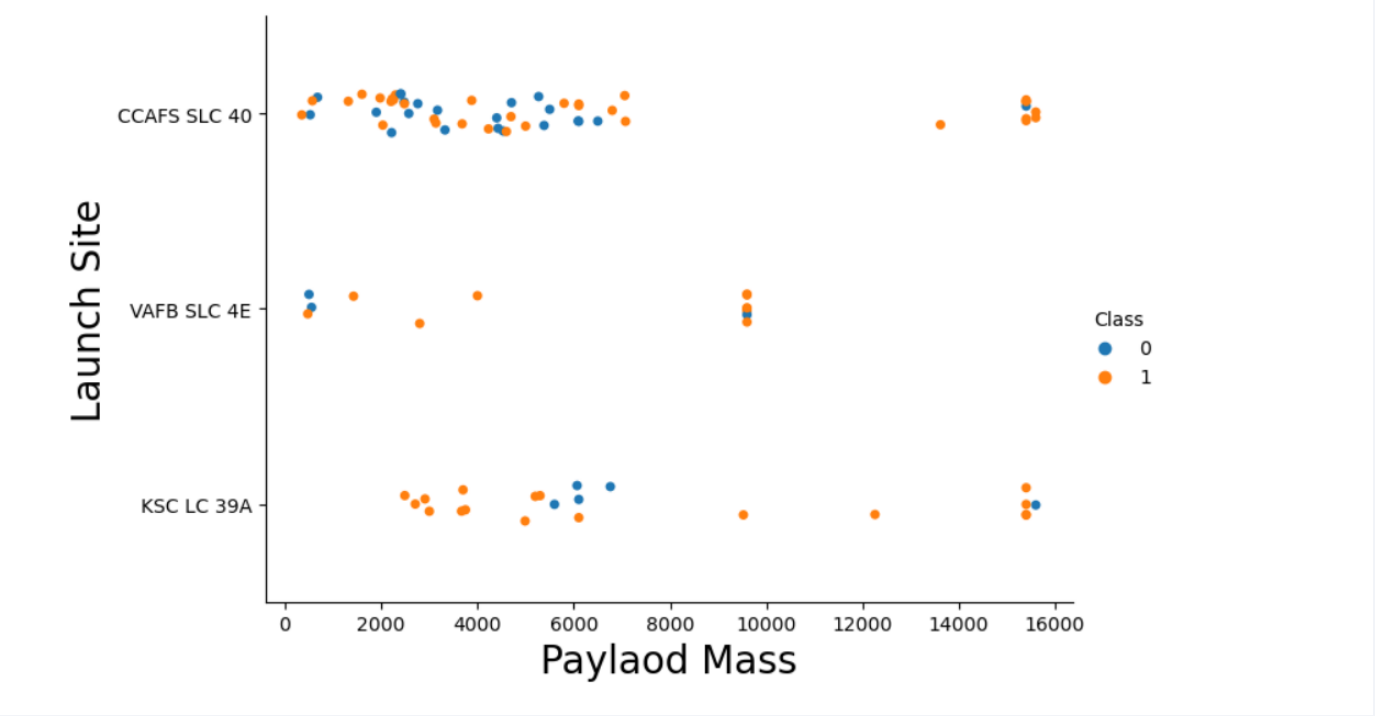
Flight Number vs. Launch Site



A scatter plot of Flight Number vs. Launch Site is displayed. It suggest that

1) Initially, the launches are implemented at CCAFS SLC40 (1-20), then from 40, CCAFS SLS40 had more launches with high success rate.
2) VAFAB SLC 4E had the launches from 20 to 60 with high success rate.
3) KSC LC 39A had launches from 22.
4) CCAFS SLC40 had more launches than other sites.

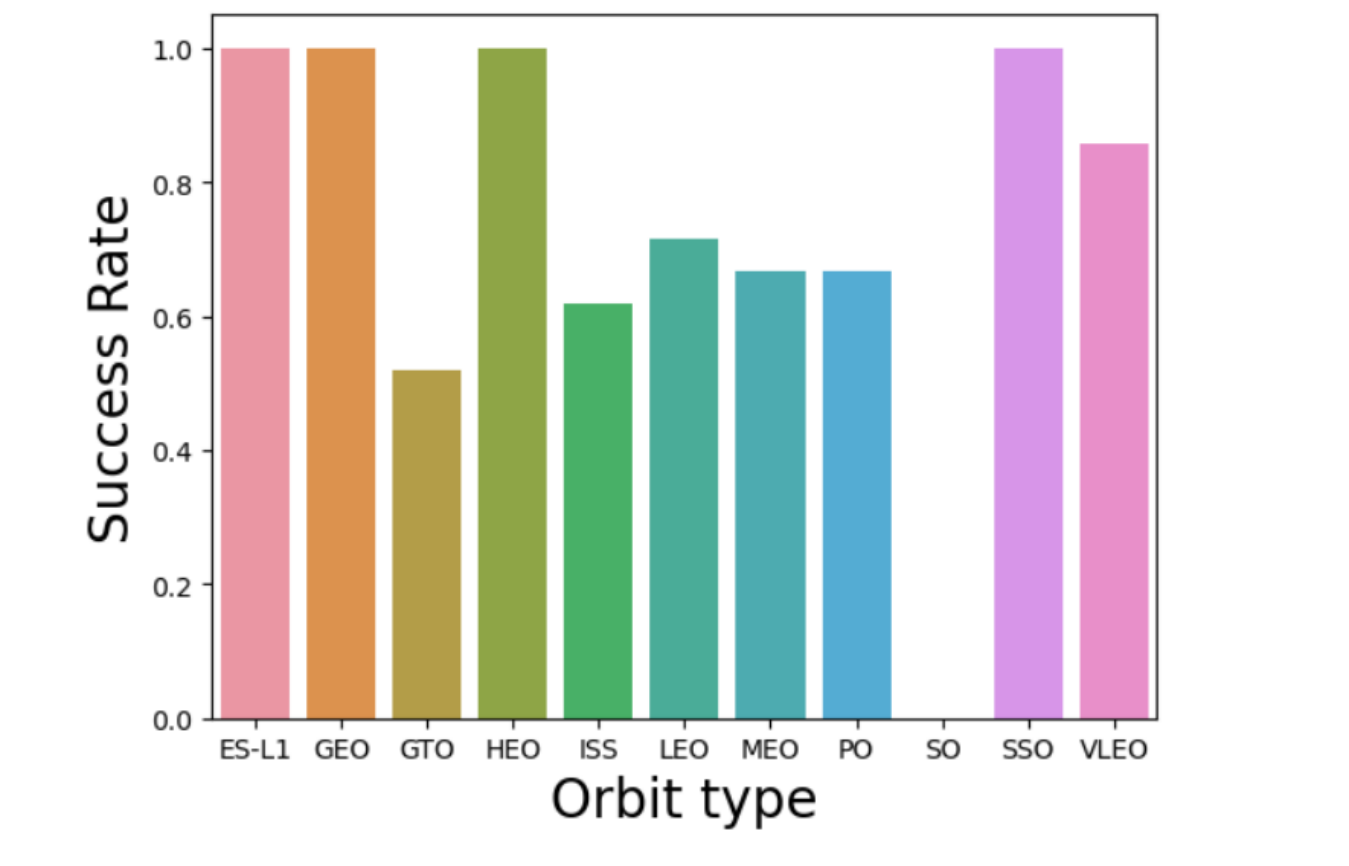
Payload vs. Launch Site



A scatter plot of Payload vs. Launch Site is displayed.

1) CCAFS SLC40 generally had the launches with small payload mass.
2) VAFAB SLC 4E also had light payload mass launches.
3) KSC LC 39A had the launces with the payload mass in a broad range.

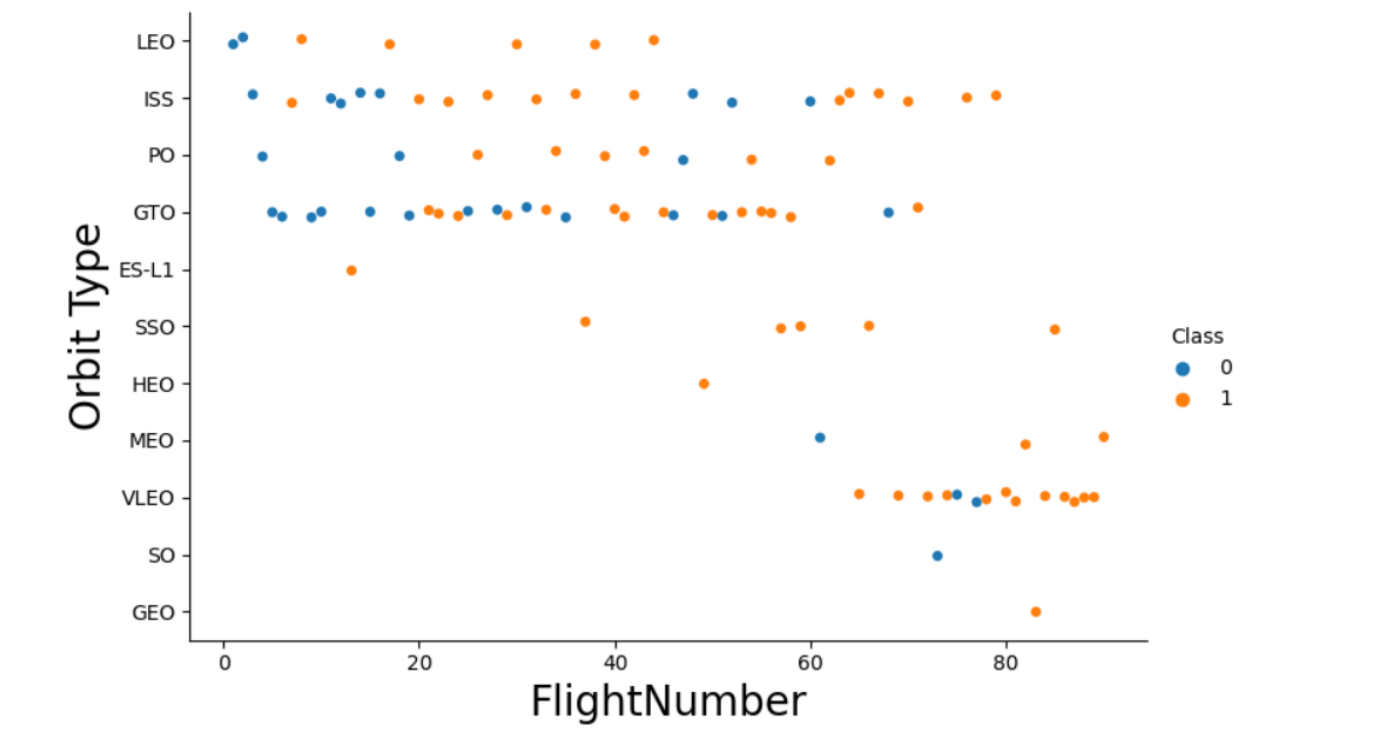
Success Rate vs. Orbit Type



A bar chart for the success rate of each orbit type is displayed.

1) Launches to ES-L1, GEO, HEO, and SSO orbits had high succuss rate.
2) GTO orbit had low success rate.

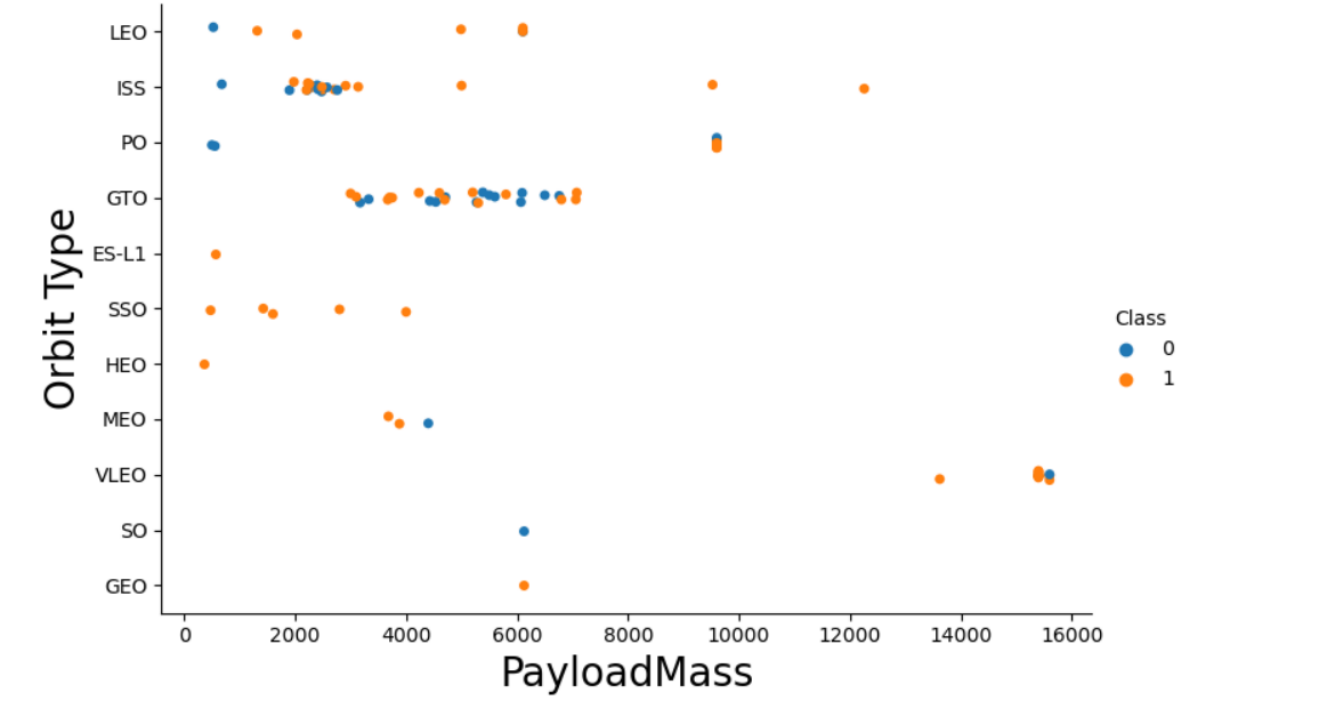
Flight Number vs. Orbit Type



A scatter point of Flight number vs. Orbit type is displayed.

The LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type

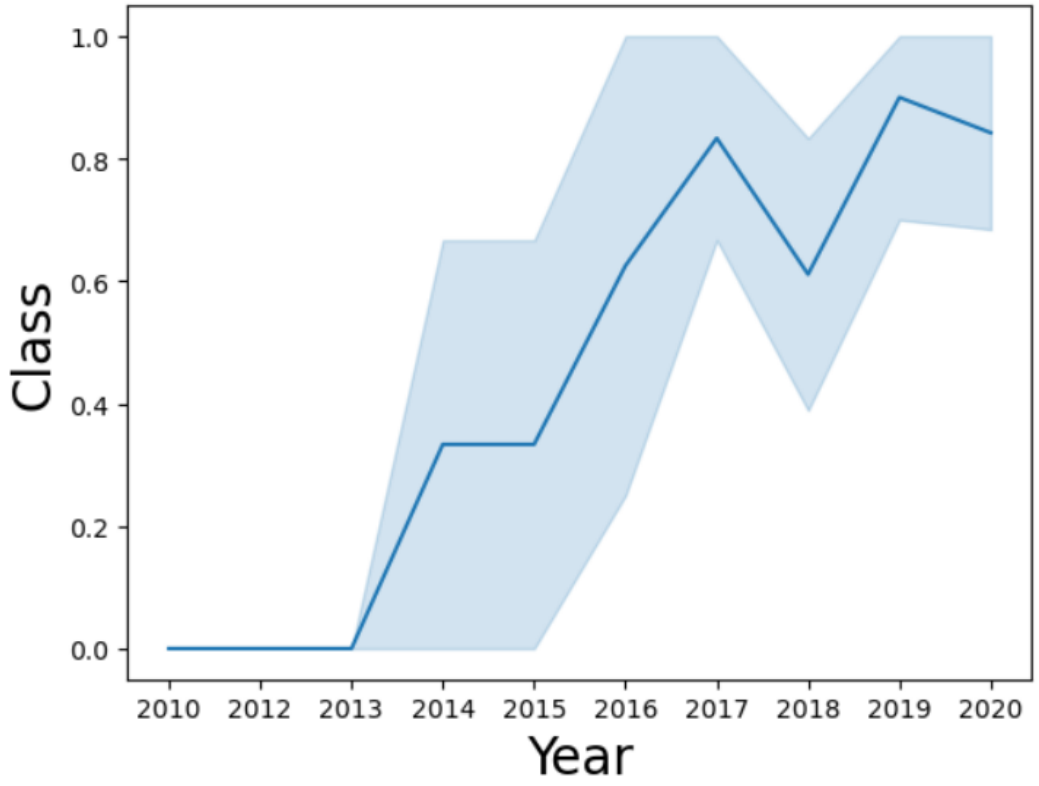


A scatter point of payload vs. orbit type is displayed.

1) With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.

2) However, for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend



A line chart of yearly average success rate is displayed.

We can observe that the success rate since 2013 kept increasing till 2020.

All Launch Site Names

```
[10]: %sql select distinct Launch_Site from SPACEXTABLE
 * sqlite:///my_data1.db
Done.
```

[10]:

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

- Find the names of the unique launch sites
- Query results are shown here. The launch sites are:

  CCAFS LC-40

  WAFB SLC-4E

  KSC LC-39A

  CCAFS SLC-40

Launch Site Names Begin with 'CCA'

Task 2

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

[11]:
```
%sql select * from SPACEXTABLE where Launch_Site like 'CCA%' limit 5
```

```
 * sqlite:///my_data1.db
Done.
```

[11]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- Find 5 records where launch sites begin with `CCA`
- The query result is displayed above that shows first 5 launches at CCAFS LC-40

Total Payload Mass

Task 3

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

```
[12]: %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer='NASA (CRS)'
```

```
 * sqlite:///my_data1.db
Done.
```

[12]:

sum(PAYLOAD_MASS__KG_)
45596

- Calculate the total payload carried by boosters from NASA
- The query result is displayed above with the total payload carried by boosters from NASA of 45596 kg.

Average Payload Mass by F9 v1.1

Task 4

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

```
[13]: %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version like 'F9 v1.1%'
```

```
 * sqlite:///my_data1.db
Done.
```

[13]:

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

- Calculate the average payload mass carried by booster version F9 v1.1
- The average payload mass carried by booster version F9 v1.1 is 2534.7 kg.

First Successful Ground Landing Date

Task 5

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

Hint:Use min function

```
[14]: %sql select min(Date) from SPACEXTABLE where Landing_Outcome='Success (ground pad)'
```

```
 * sqlite:///my_data1.db
Done.
```

[14]:

min(Date)
2015-12-22

- Find the date of the first successful landing outcome on ground pad
- The date of the first successful landing outcome on ground pad was December 22, 2015.

Successful Drone Ship Landing with Payload between 4000 and 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

```
[15]: %sql select Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTABLE where PAYLOAD_MASS__KG_ between 4000 and 6000
```

```
 * sqlite:///my_data1.db
Done.
```

[15]:

Booster_Version	PAYLOAD_MASS__KG_
F9 v1.1	4535
F9 v1.1 B1011	4428
F9 v1.1 B1014	4159
F9 v1.1 B1016	4707
F9 FT B1020	5271
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1030	5600
F9 FT B1021.2	5300
F9 FT B1032.1	5300
F9 B4 B1040.1	4990
F9 FT B1031.2	5200
F9 B4 B1043.1	5000
F9 FT B1032.2	4230
F9 B4 B1040.2	5384
F9 B5 B1046.2	5800
F9 B5 B1047.2	5300
F9 B5 B1046.3	4000
F9 B5B1054	4400
F9 B5 B1048.3	4850
F9 B5 B1051.2	4200
F9 B5B1060.1	4311
F9 B5 B1058.2	5500
F9 B5B1062.1	4311

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

- The query result is shown here.

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

```
23]: %sql select count(*) from SPACEXTABLE where (Landing_Outcome like 'Success%') or (Landing_Outcome like 'Failure%')
 * sqlite:///my_data1.db
Done.
23]: count(*)
71
```

- Calculate the total number of successful and failure mission outcomes
- The total number of successful and failure mission outcomes is 71.

Boosters Carried Maximum Payload

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

```
[18]: %sql select Booster_Version, PAYLOAD_MASS__KG_ from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)
```

```
 * sqlite:///my_data1.db
Done.
```

[18]:

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

- List the names of the booster which have carried the maximum payload mass
- A list of booster names with the maximum payload mass of 15600 kg is shown above.

2015 Launch Records

```
[20]: %sql select substr(Date, 6, 2), Landing_Outcome, Booster_Version, Launch_Site from SPACEXTABLE \
      where Landing_Outcome='Failure (drone ship)' and substr(Date,0,5) = '2015'
```

```
 * sqlite:///my_data1.db
Done.
```

[20]:

substr(Date, 6, 2)	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- In the of 2015, the failed landing outcomes in drone ship, their booster versions, and launch site names are shown above.

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

```
[33]: %%sql

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

```
 * sqlite:///my_data1.db
Done.
```

[33]:

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

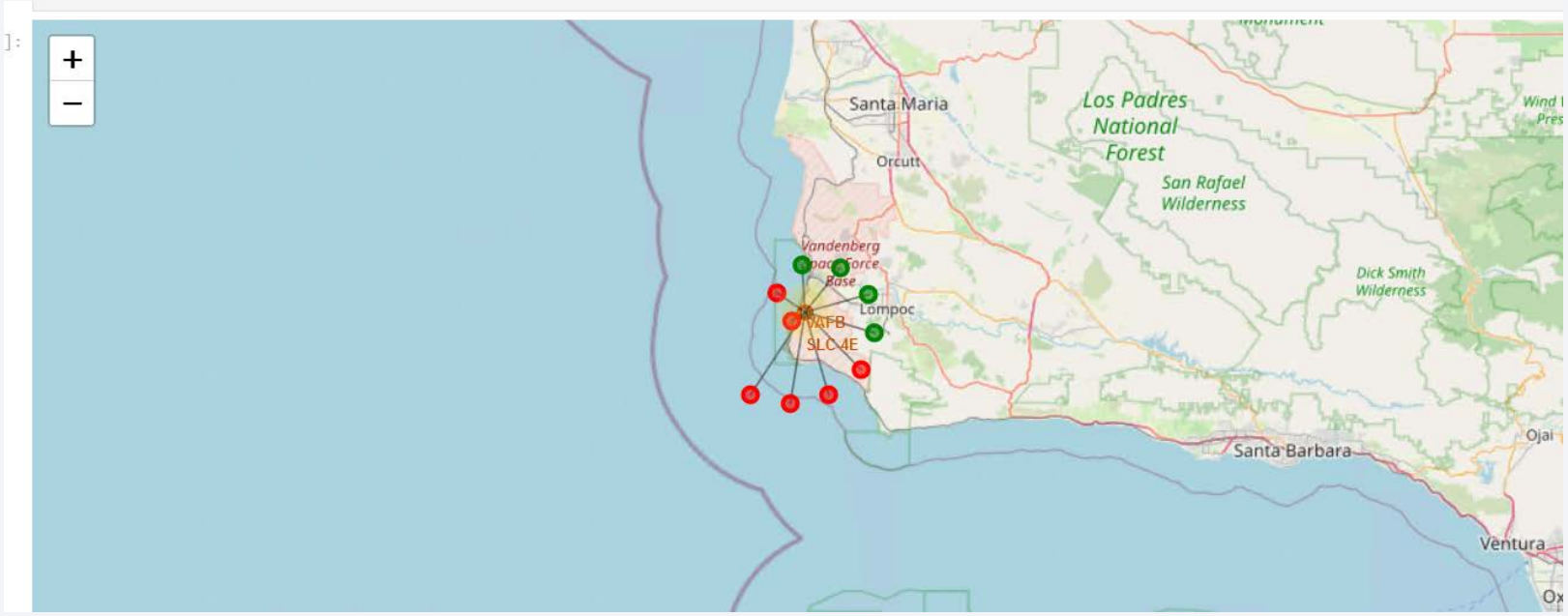
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order
- The count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order is shown here.


Section 3
Launch Sites
Proximities Analysis

Launch Sites shown in map

- The generated folium map with all launch sites' location markers

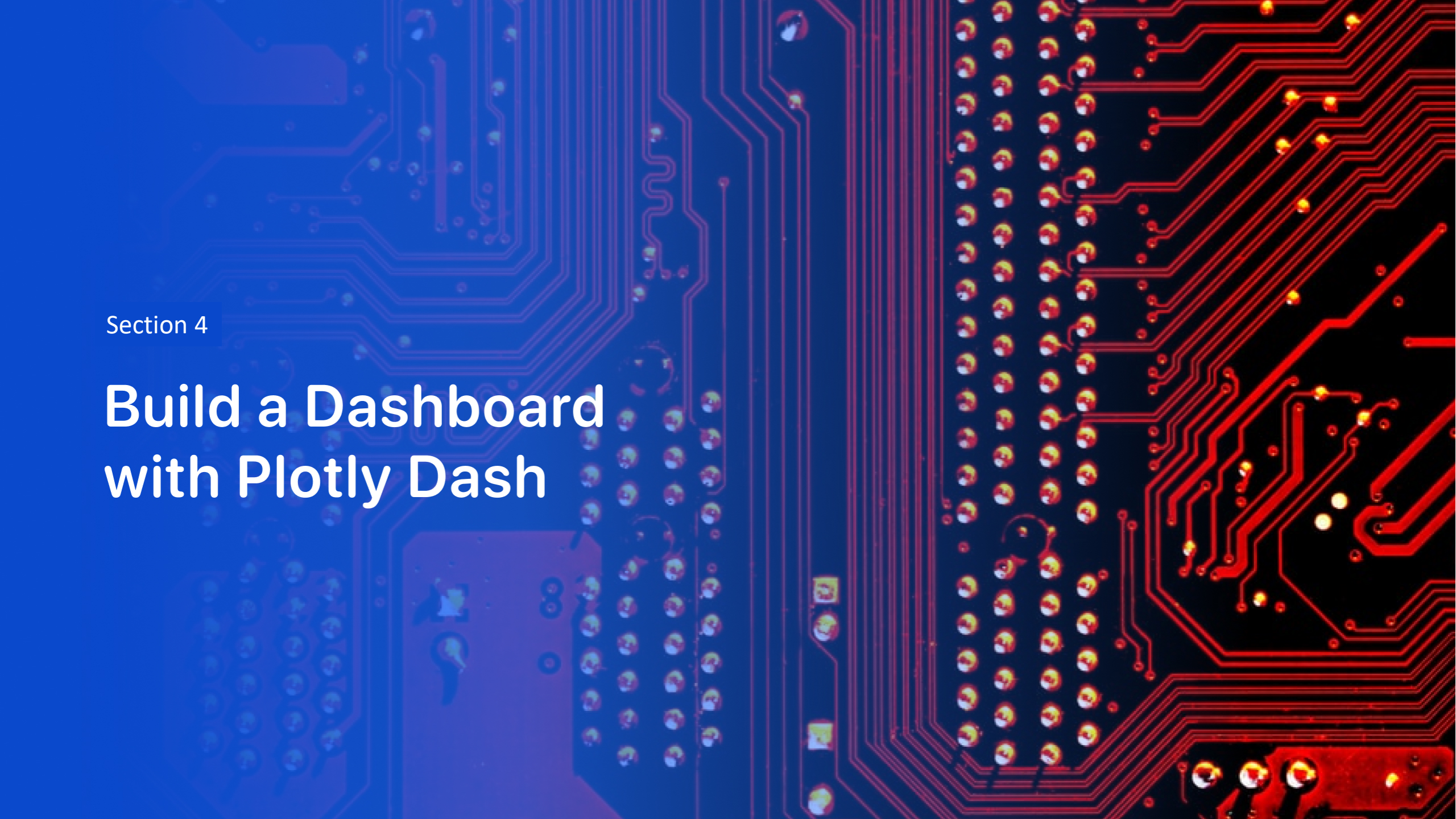
Success rate at VAFB SLC-4E shown in map



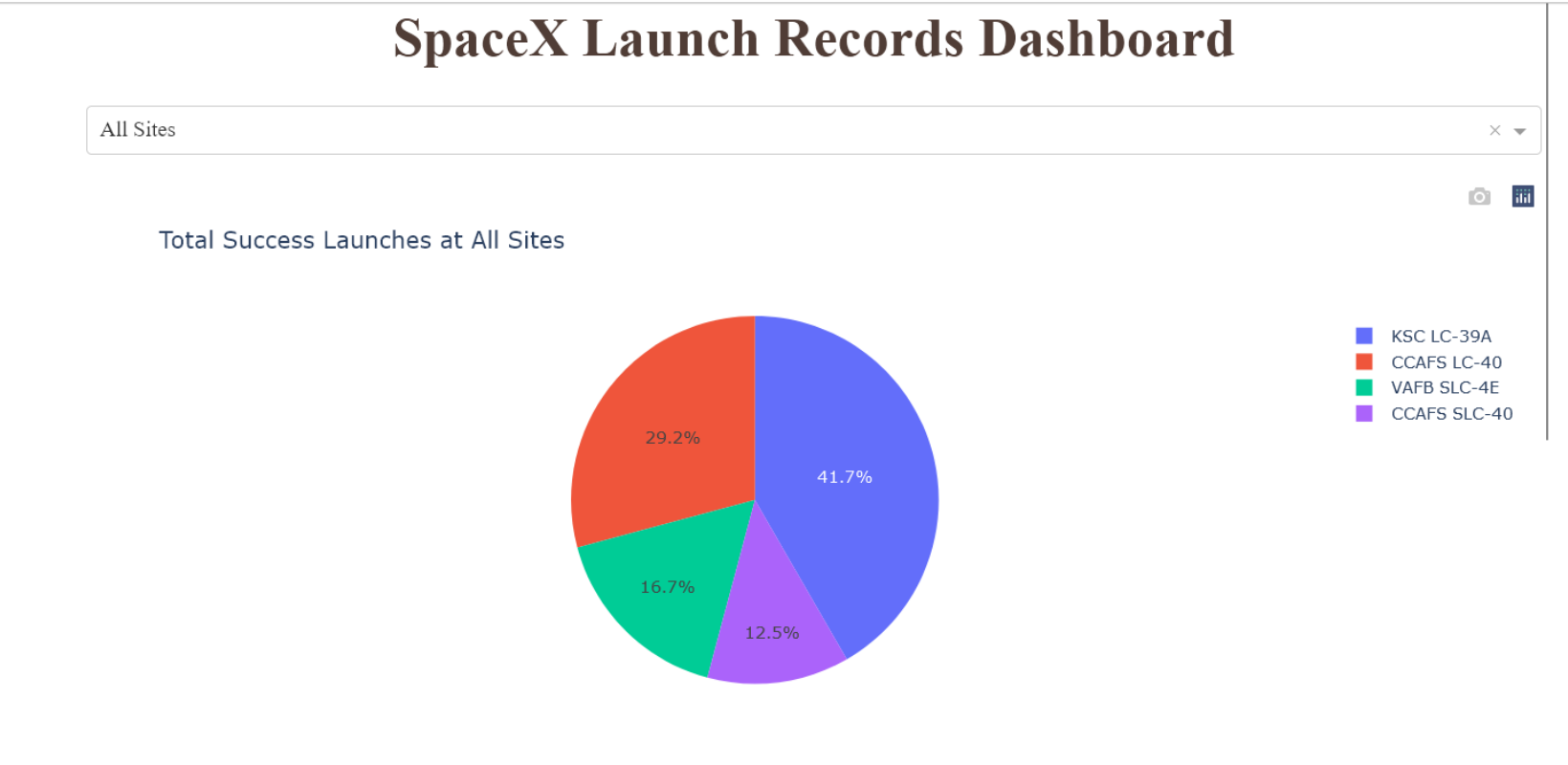
- The folium map showing the color-labeled launch outcomes

Distance between nearest beach and VAFB SLC-4E

- The generated folium map showing the selected launch site (WAFB SLC-4E) to its proximities of a beach with distance (1.32 km) calculated and displayed.

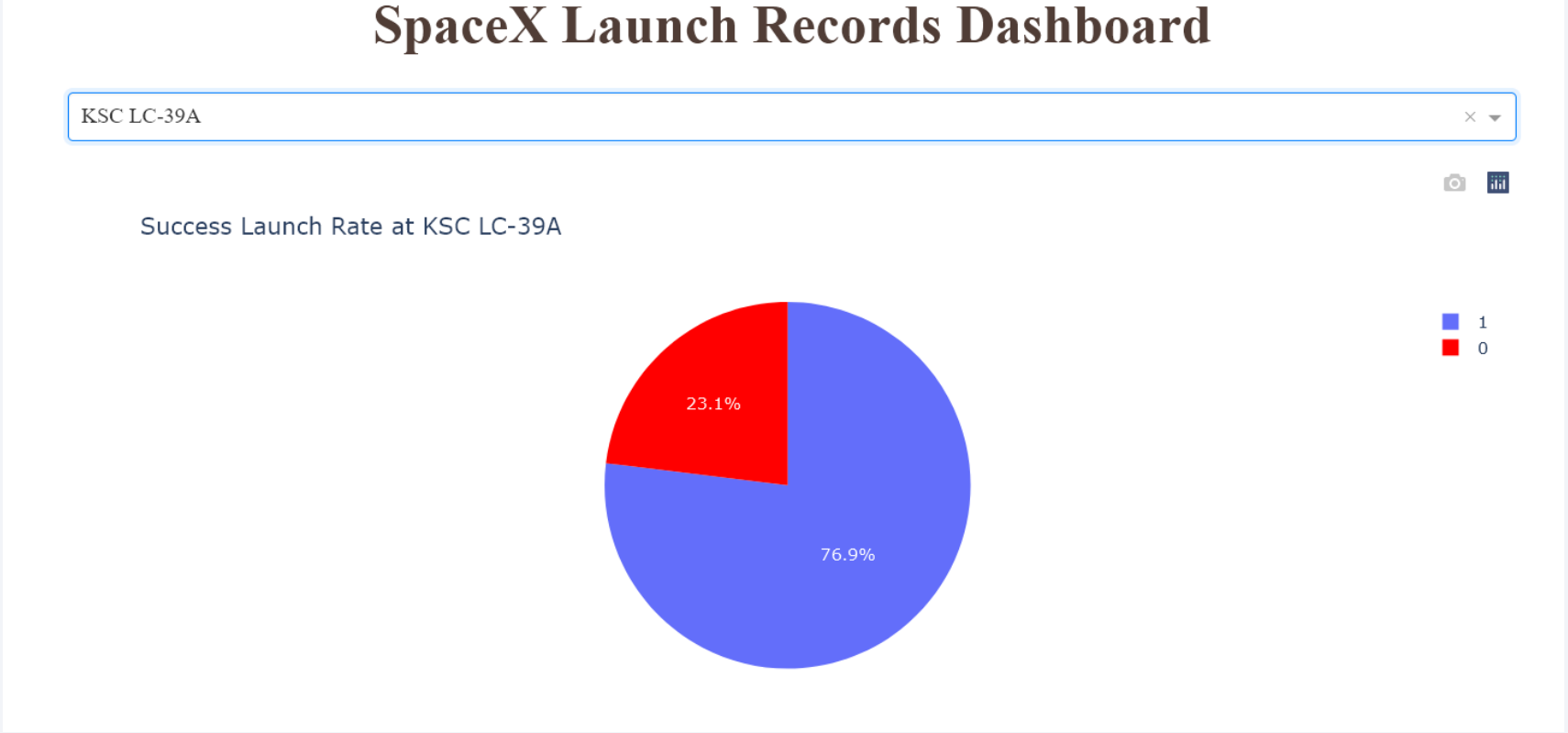

Section 4
Build a Dashboard with Plotly Dash

Dashboard of launch success count for all sites



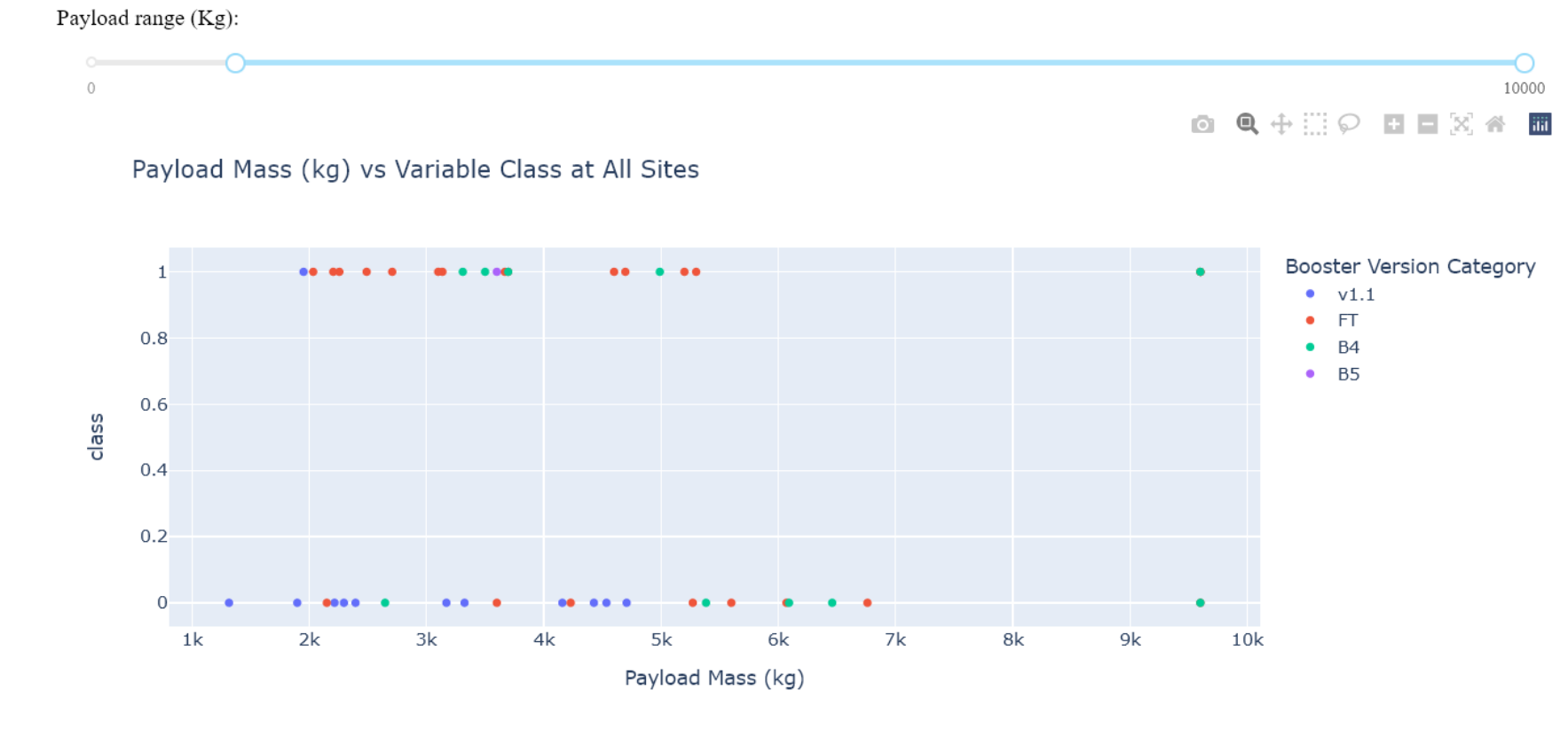
- KSC LC-39A had the highest success rate of 41.7%.
- CCAFS SLC-40 had the lowest success rate of 12.5%.

Dashboard of launch success count for KSC LC-9A



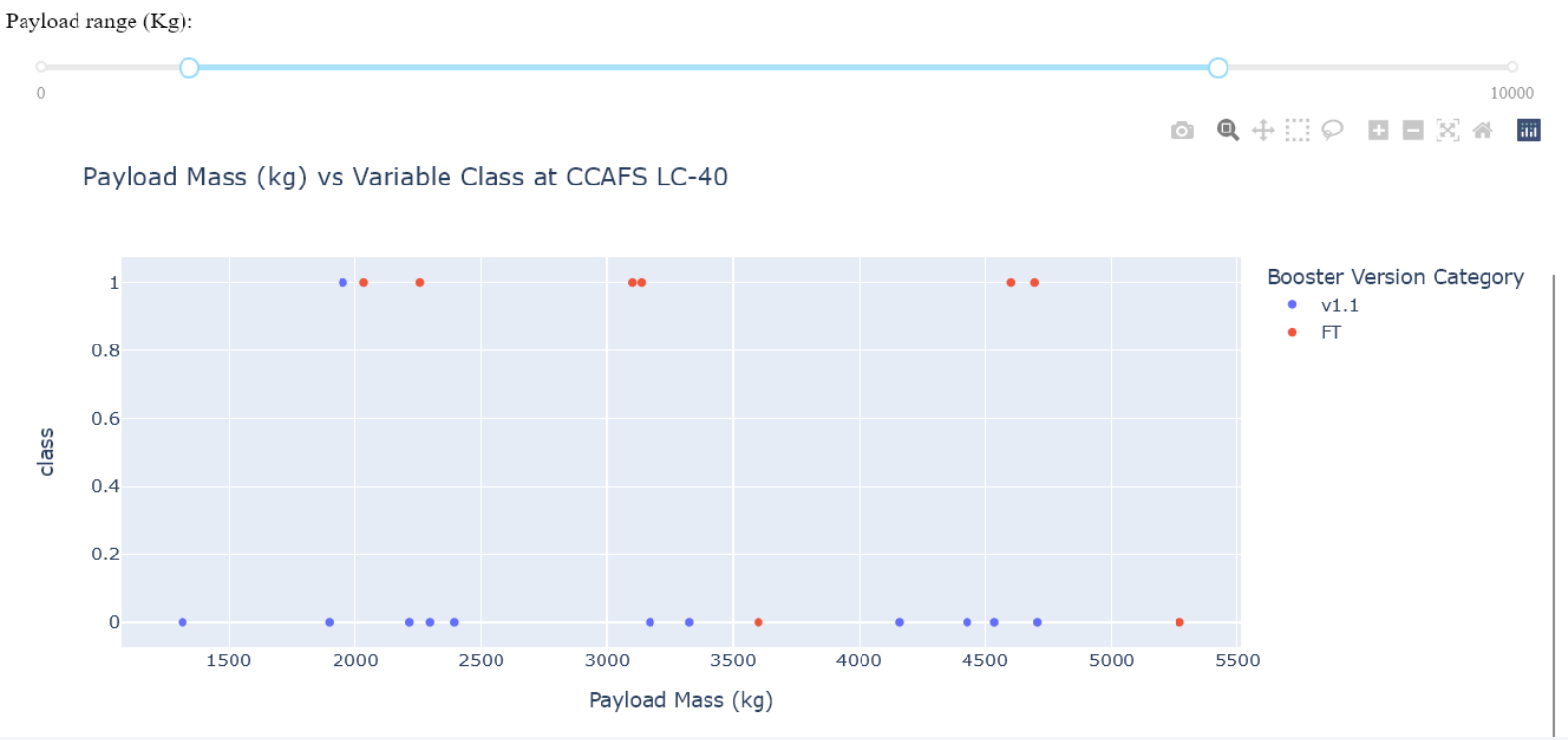
- KSC LC-9A had a success rate of 76.9%.

Dashboard of Payload vs. Launch Outcome for all sites



- The booster version of FT had the largest success rate.
- The payload mass between 2K and 5.5K had the high suc

Dashboard of Payload vs. Launch Outcome at CCAFS LC-40

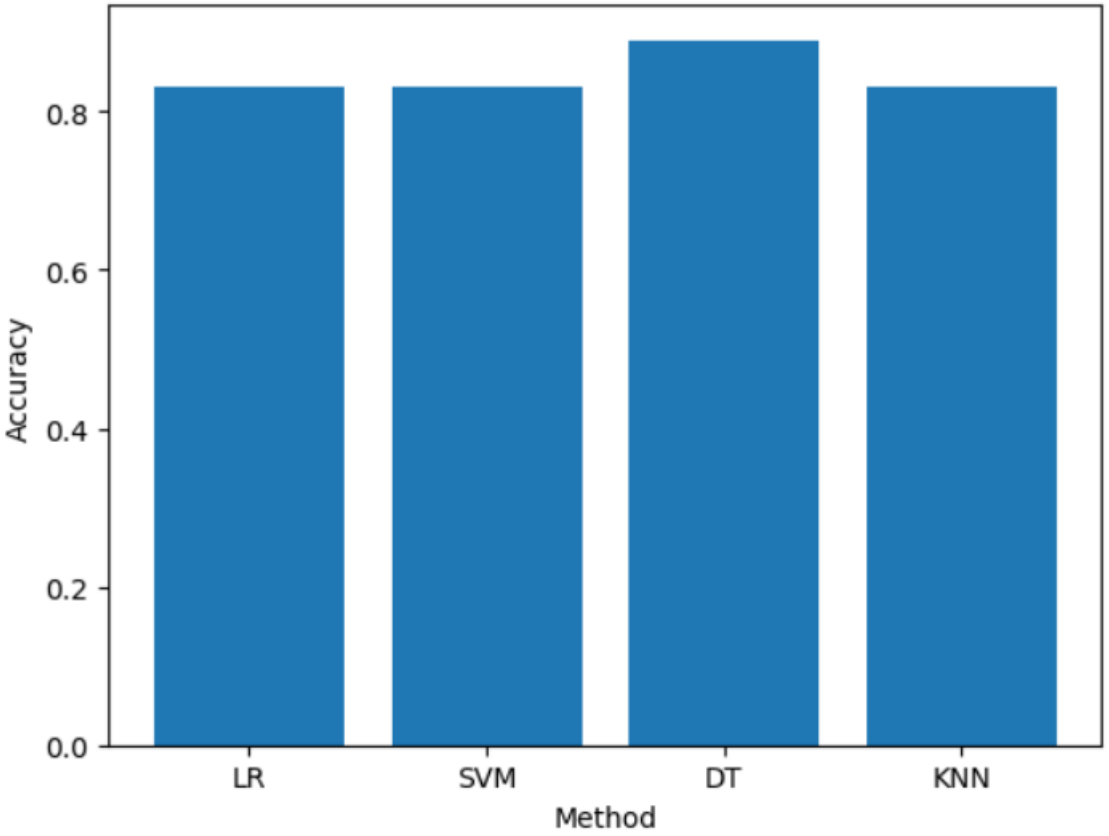


- The booster version of FT had the largest success rate.

Section 5

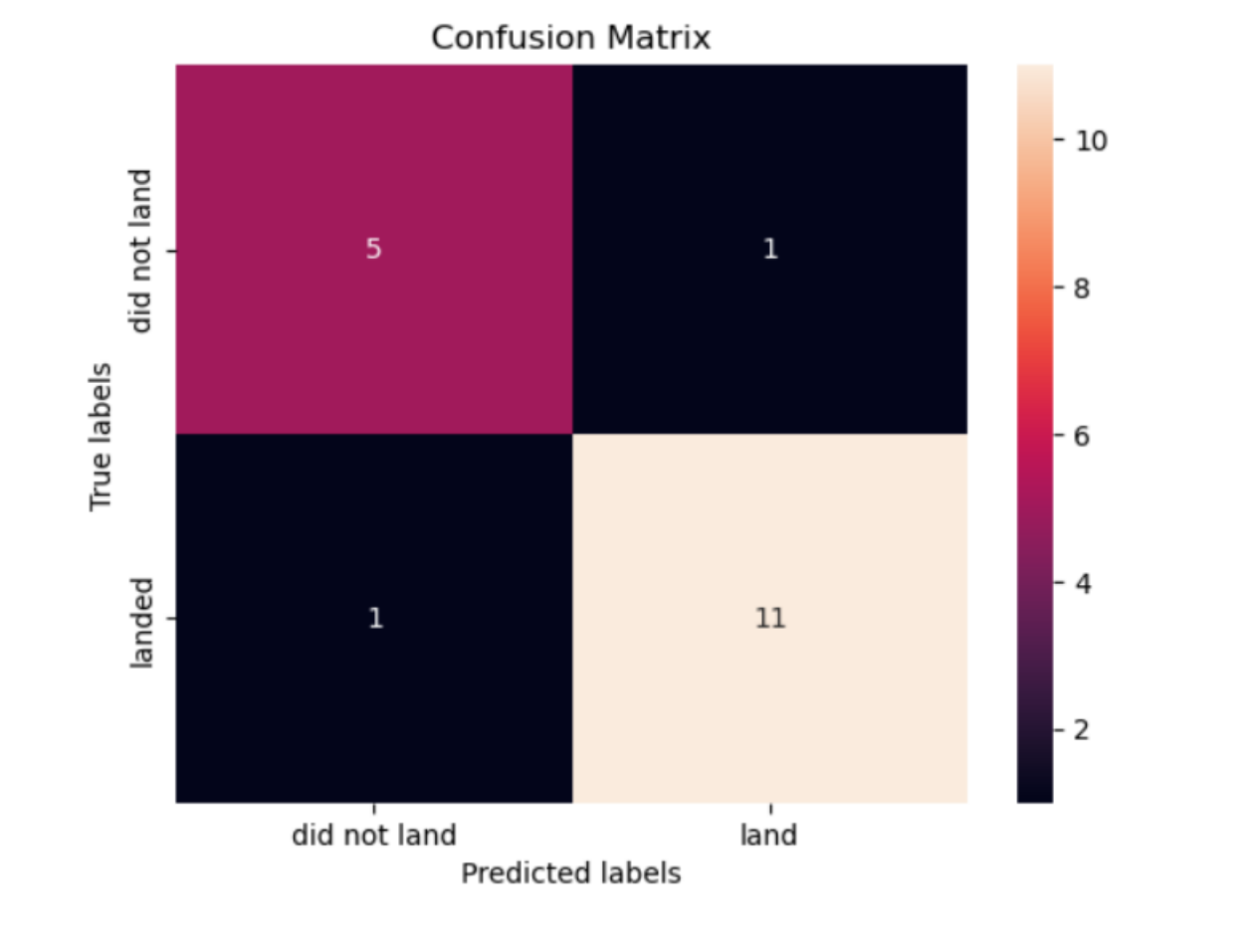
Predictive Analysis (Classification)

Classification Accuracy



- The Decision Tree model has the highest classification accuracy

Confusion Matrix



- For the “didn’t land” cases, 5 out of 6 are well predicted.
- For the “landed” cases, 11 out of 12 are well predicted.

Conclusions

- 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%.
- For the VAFB-SLC launchsite there are no rockets launched for heavy payload mass (greater than 10000).
- The success rate since 2013 kept increasing till 2020.
- The Decision Tree has the highest accuracy (R2) of 0.89 for the test dataset in predicting the outcome of a launch.

Appendix

- All the python notebooks for this project are uploaded on:

https://github.com/panj1963/Applied_Data_Science_Capstone

Thank you!