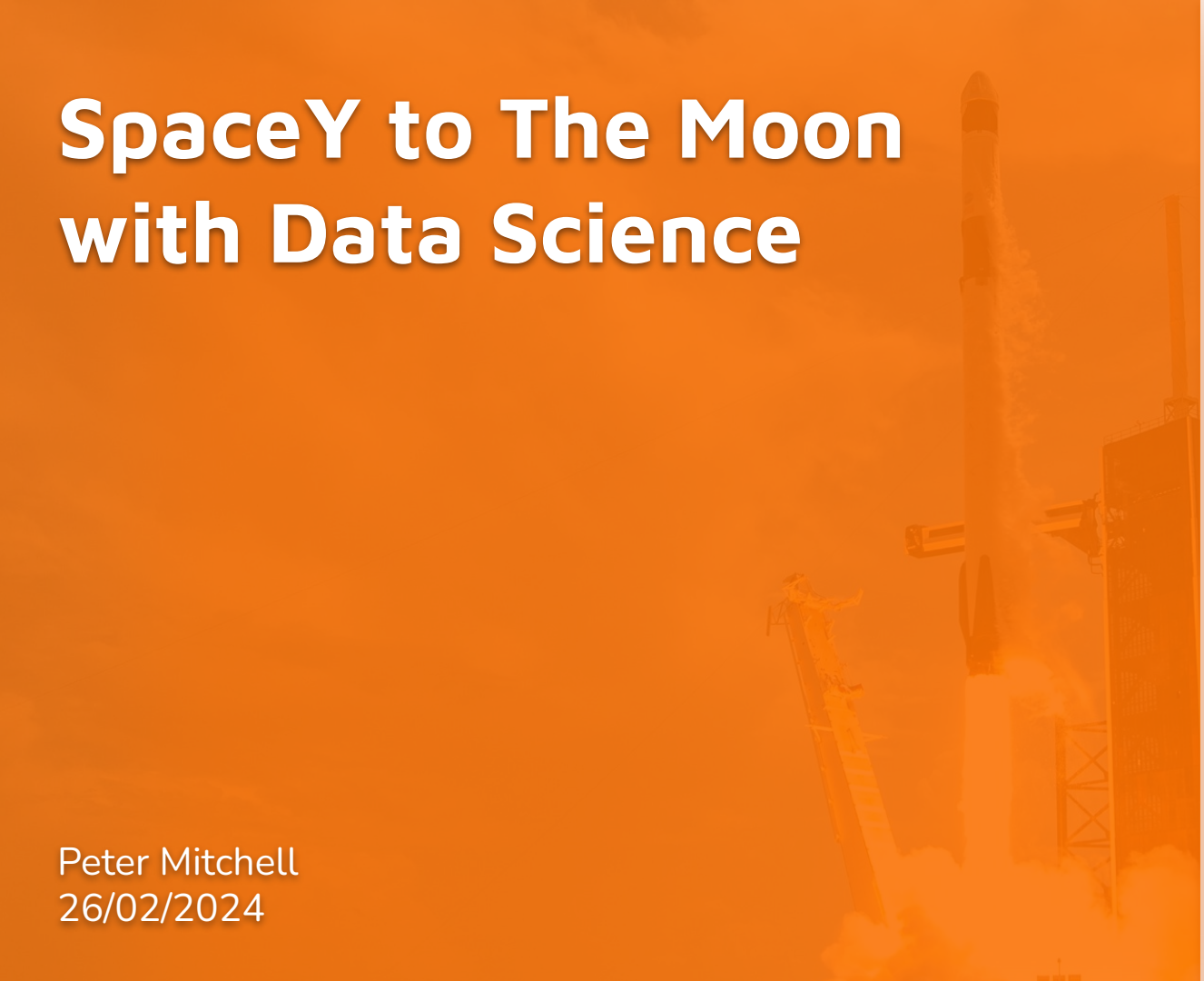
SpaceY to The Moon with Data Science



Peter Mitchell
26/02/2024

IBM Developer
SKILLS NETWORK

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results + Analysis
- Conclusion

EXECUTIVE SUMMARY

The data was analysed through the following methodologies:

- Web scraping and SpaceX API Data collection
- Exploratory Data Analysis (EDA); Data wrangling, visualisation and dashboards
- Machine Learning

Summary of Results

- Publicly available data was filtered into actionable insights
- EDA identified features for successful launches
- Machine Learning predicted model required to increase success and profitability of launches

Introduction

SpaceY aims to compete with SpaceX.

In order to do so Mission Objectives are as follows:

1. SpaceY needs to be able to estimate the cost for any given launch.

2. Factoring in the predictability of a successful first stage rocket launch.

3. Find the best site to launch from to support points 1 and 2.

METHODOLOGY

Data Collection:

- Data extracted from Space X API: (https://api.spacexdata.com/v4/rockets)
- Data extracted from Web Scraping: (https:77en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)

Exploratory Data Analysis:

- Data Wrangling (data quality was enriched by structuring and transforming raw data into a 9)
- Data Visualisation (designed easy-to-communicate graphics representing complex qualitative and quantitative data)
- SQL (structured query language used to search data frame for further insights)

Interactive Map/Dashboard Generation:

- Dashboards using Plotly Dash and Interactive Maps made from Folium

Machine Learning:

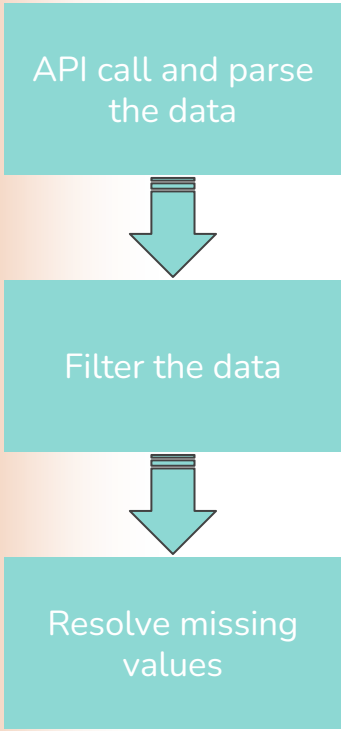
- Predictive analysis using classification models

Data Collection - SpaceX API

- SpaceX provides a public API where data can be obtained
- API calls were made to retrieve data on:

  BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

- Source code: https://github.com/Peter-mitch1/Data-Science-Capstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb

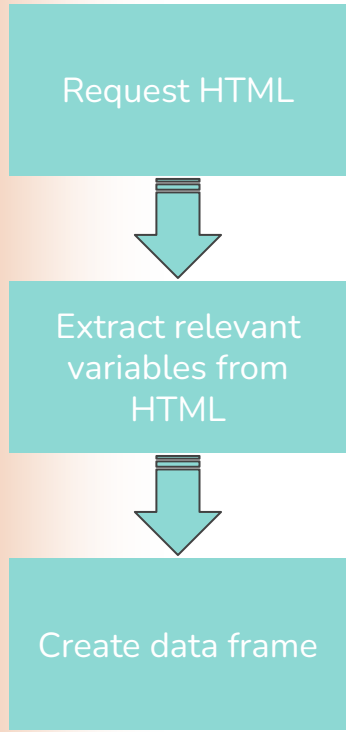


Data Collection - Web Scraping

- SpaceX data can be obtained from Wikipedia:https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- HTML request was made and the relevant variables were extracted and placed into a data frame:

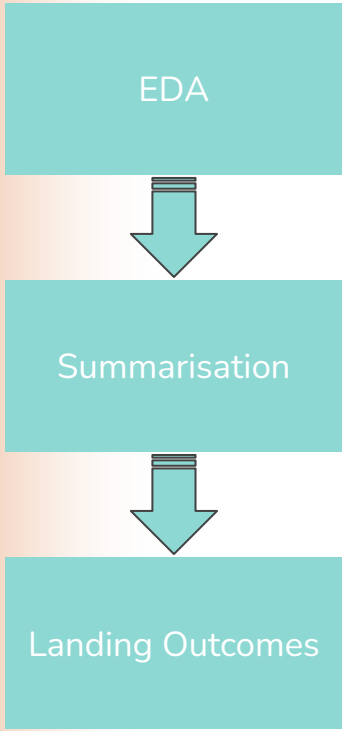
	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	1	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0004.1	Failure	4 June 2010	18:45
2	2	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0005.1	No attempt\n	8 December 2010	15:43
3	3	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.0B0006.1	No attempt	22 May 2012	07:44
4	4	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.0B0007.1	No attempt\n	8 October 2012	00:35
...	...	...	...	...	...	...	...	...	...	...	...
117	117	KSC	Starlink	~14,000 kg	LEO	SpaceX	Success\n	F9 B5B1058.8	Success	9 May 2021	06:42
118	118	CCSFS	Starlink	15,600 kg	LEO	SpaceX	Success\n	F9 B5B1063.2	Success	15 May 2021	22:56
119	119	KSC	SpaceX CRS-22	3,328 kg	LEO	NASA	Success\n	F9 B5B1067.1	Success	26 May 2021	18:59
120	120	CCSFS	SXM-8	7,000 kg	GTO	Sirius XM	Success\n	F9 B5	Success	3 June 2021	17:29
121	121	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	6 June 2021	04:26

- Source code: https://github.com/Peter-mitch1/Data-Science-Capstone/blob/main/jupyter-labs-webscraping.ipynb



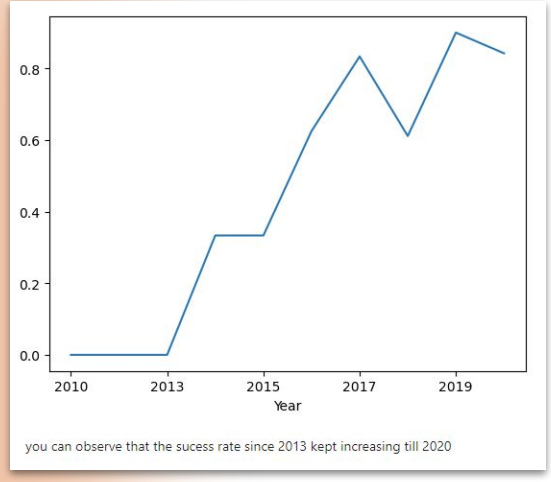
EDA - Data Wrangling

- Exploratory Data Analysis was utilised to wrangle data and find a percentage based answer to the success rate of Falcon 9s first stage landing
- The Answer 66%
- Source code: https://github.com/Peter-mitch1/Data-Science-Capstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

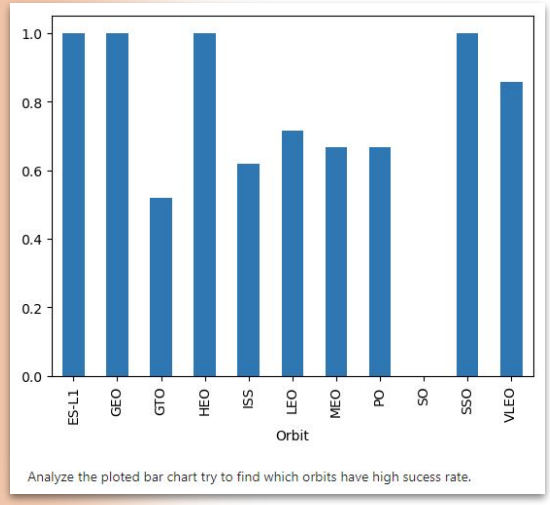


EDA - Data Visualisation

- To explore the data further and make insights actionable it was important to draw relationships through visual insights. As such barplots, line graphs and scatter plots were used to help visualise the data and derive further important key parameters and metrics.
- Source code: https://github.com/Peter-mitch1/Data-Science-Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb



you can observe that the sucess rate since 2013 kept increasing till 2020



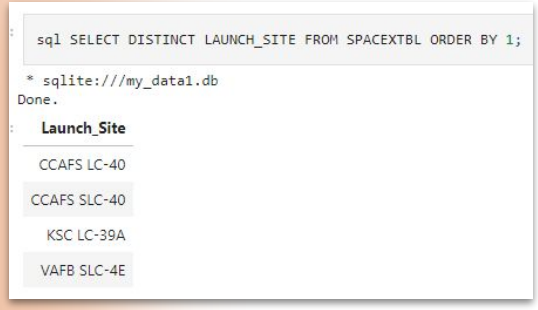
Analyze the ploted bar chart try to find which orbits have high sucess rate.

EDA - SQL

SQL (Structured Query Language) used to gather further insights:

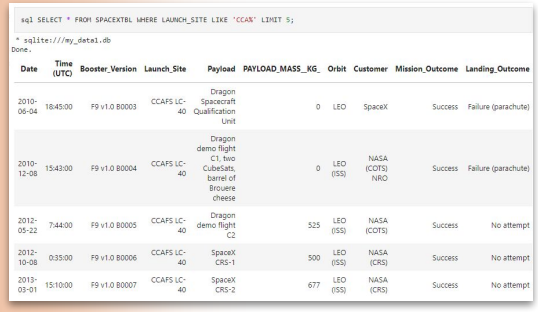
```
sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;

sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;

sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE PAYLOAD  LIKE '%CRS%';

sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';

sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (ground pad)';

sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND LANDING_OUTCOME = 'Success (drone ship)';

sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;

sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL) ORDER BY BOOSTER_VERSION;

sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Failure (drone ship)' AND DATE(2015);

sql SELECT LANDING_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING_OUTCOME ORDER BY QTY DESC;
```

Source code:
https://github.com/Peter-mitch1/Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb



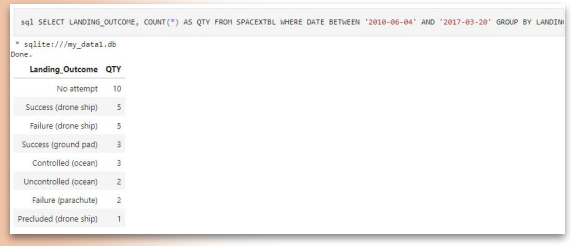
```
sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;
 * sqlite:///my_data1.db
Done.
```

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



```
sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
 * sqlite:///my_data1.db
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	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



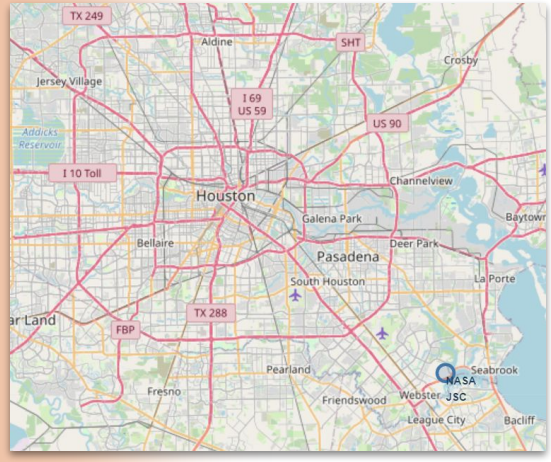
```
sql SELECT LANDING_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	QTY
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

Interactive Map

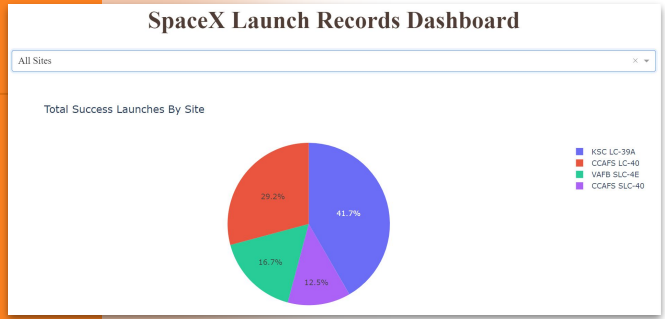
Using Folium Maps an interactive map was generated

- Circles indicate specific coordinates in this instance the surrounding base of operations of a launch site such as NASA Johnson Space Center.
- Markers indicate launch sites.
- Marker clusters indicate grouped events. Such as launches in a launch site.
- Lines display distance.



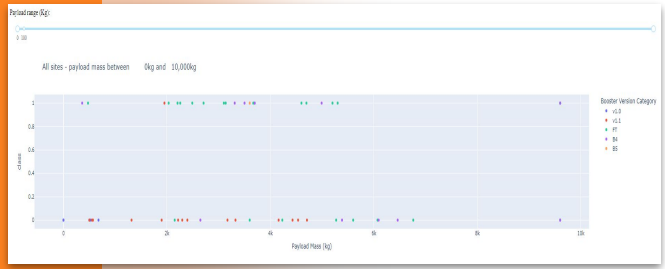
Source code:
https://github.com/Peter-mitch1/Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Dashboard Generation



Using Plotly an interactive Dashboard was generated. Dashboards make information tactile and more easily comparative.

- Graphs and plots were developed to visualise data i.e. Success of Launches by Site, Payload Range vs Success.

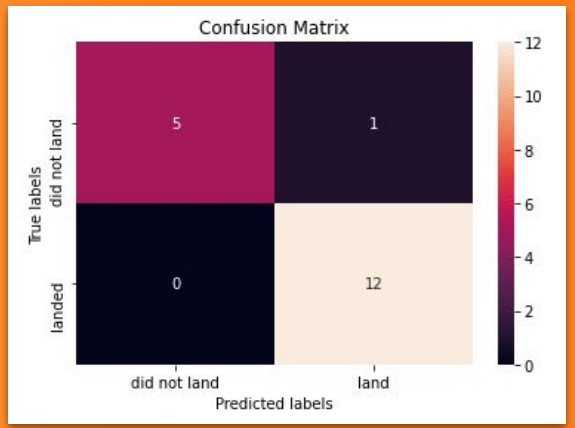


Source code: https://github.com/Peter-mitch1/Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Machine Learning - Predictive Analysis

Compare four different classification models to find the best fit:

- Logistic regression, support vector machine, k nearest neighbours, decision tree.
- Source code: https://github.com/tflores/applied-data-science-capstone/blob/master/Machine%20Learning%20Prediction.ipynb



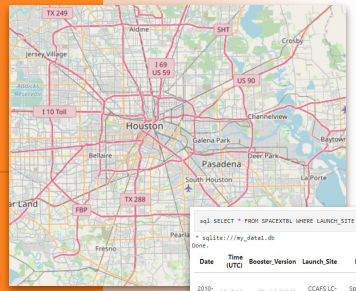
Standardise and Prepare Data

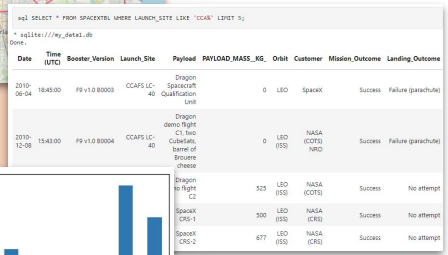
Test Models with combinations of hyperparameters

Compare Results

Results + Analysis

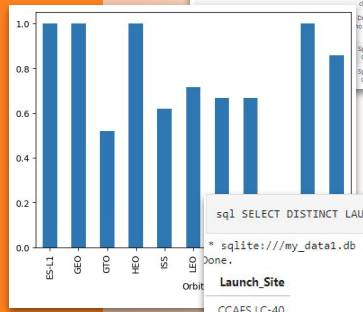
In the following slides all results will be presented with light analysis provided to help enhance the data





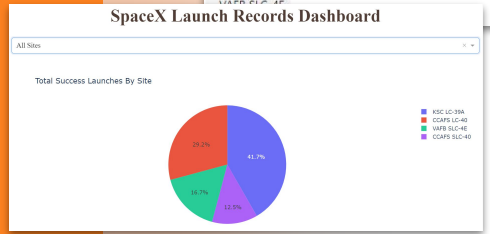
```
sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCAS' LIMIT 5;
 * sqlite:///my_data1.db
Done.
```

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)
				Dragon to flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
				SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
				SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt




```
sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;
 * sqlite:///my_data1.db
Done.
```

Launch_Site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A



Results - EDA SQL

```
sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;
```

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

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

FIRST_SUCCESS_GP
2015-12-22

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 SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

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

Rank the count of landing out
20, in descending order.

```
sql SELECT LANDING_OUTCOME
```

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

Landing_Outcome	QTY
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

```
sql SELECT BOOSTER_VERSION,
```

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

Booster_Version	Launch_Site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40
F9 v1.1 B1017	VAFB SLC-4E
F9 FT B1020	CCAFS LC-40
F9 FT B1024	CCAFS LC-40

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 SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND LANDING_OUTCOME = 'Succ
```

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

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

List the total number of successful and failure mission outcomes

```
sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;
```

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

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

Results - EDA SQL

Task 2

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

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

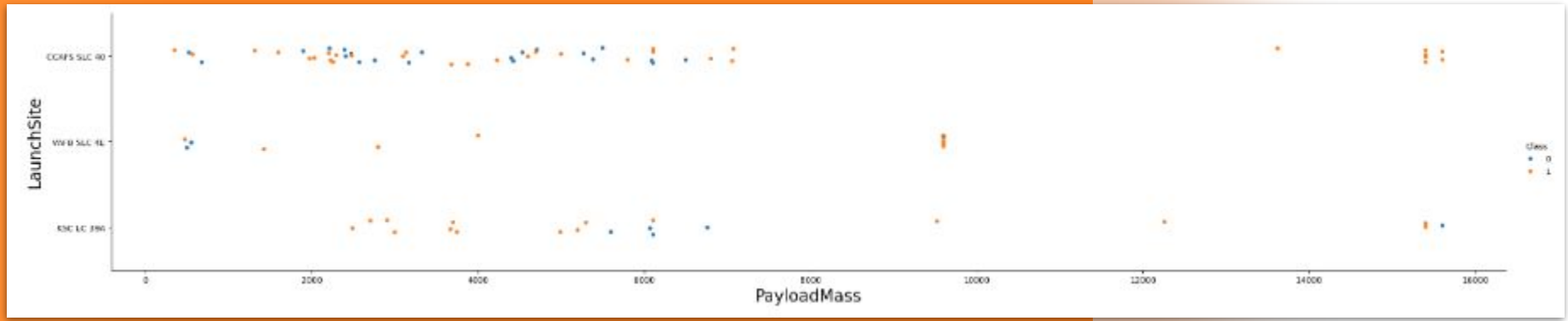
```
 * sqlite:///my_data1.db
Done.
```

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	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

Synopsis:

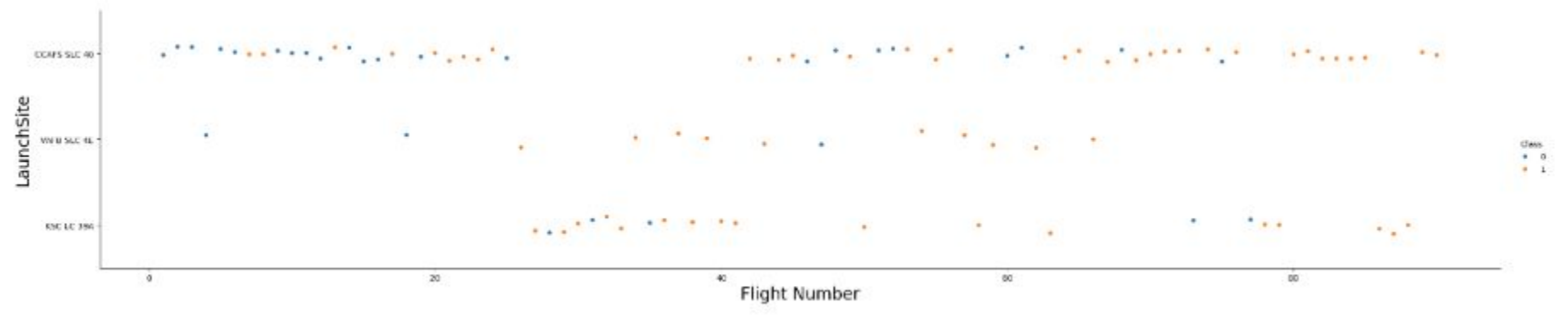
- SpaceX uses 4 different launch sites
- The average payload of F9v1.1 = 2928.4 kg
- First successful landing was in 2015
- Landing outcomes became better as years went on

Results EDA - Launch Site vs Payload Mass



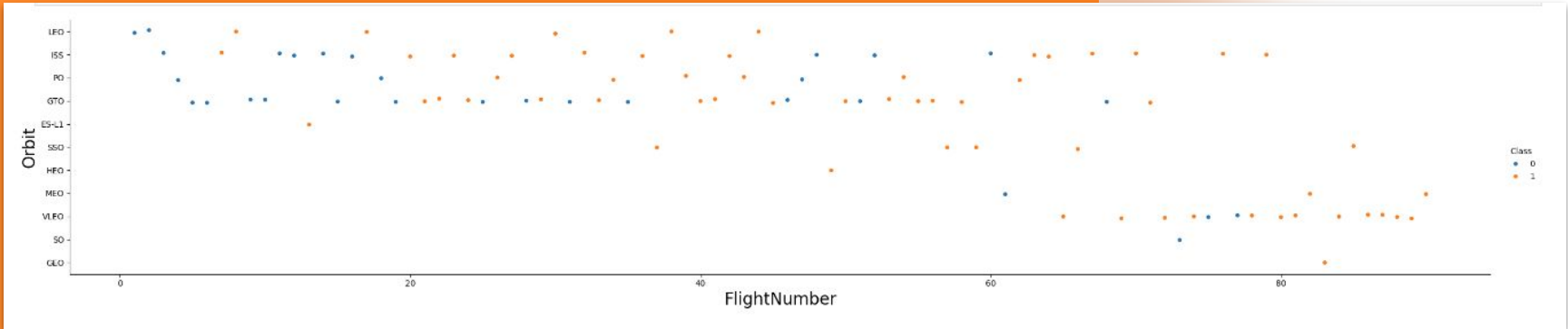
- Just under 10 tonne payloads from launch site Vandenberg Space Launch Complex 4 display a more consistent success rate than any other payload and launch site under 8 tonne.
- Also any payload above 12 tonne from Cape Canaveral Space Launch Complex 40 have a 100% success rate
- Source code: https://github.com/Peter-mitch1/Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Results EDA - Launch Site vs Flight Number



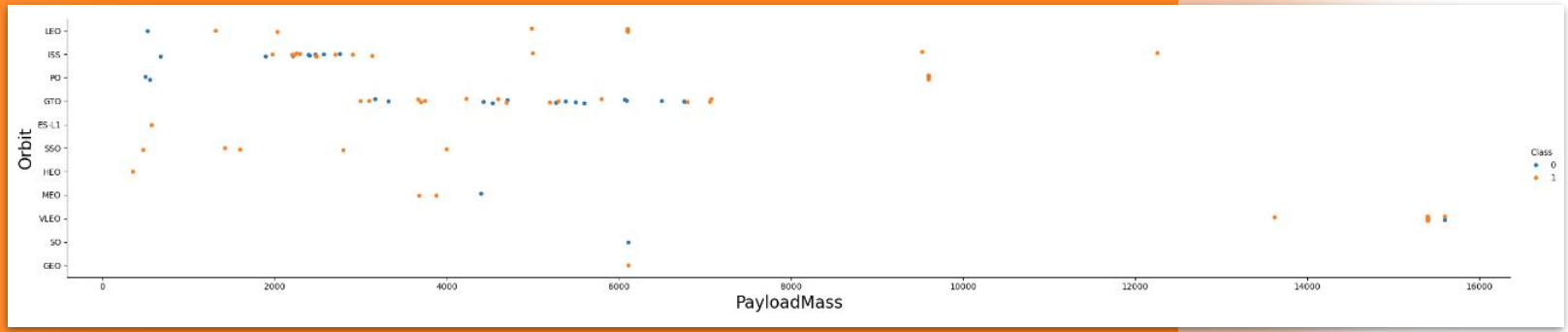
- Cape Canaveral Space Launch Complex 40(CCAF5 SLC 40) has had the most recent successful launches.
- 23.07% of launches have failed from Vandenberg Space Launch Complex 4 (VAFB SLC 4)
- 22.27% of launches have failed from Kennedy Space Center Launch Complex 39A
- Source code: https://github.com/Peter-mitch1/Data-Science-Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

Results EDA - Flight Number vs Orbit Type



- Success rate increased over time in all orbits
- VLEO appears to be a favourable orbit with an increase in frequency being noted recently
- Source code: https://github.com/Peter-mitch1/Data-Science-Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

Results EDA - Payload vs Orbit



- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- Source code: https://github.com/Peter-mitch1/Data-Science-Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

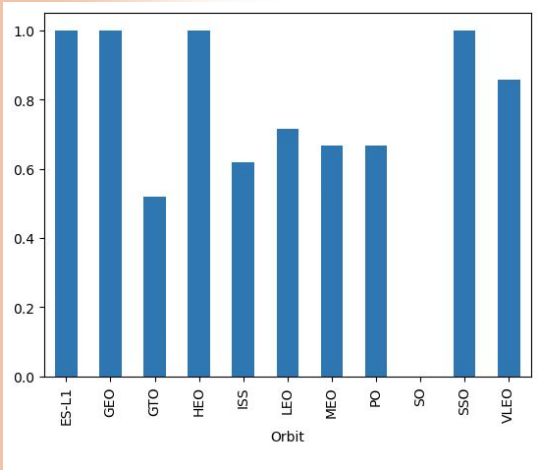
Results EDA - Success Rate vs Orbit

Best success rate happens in:

- ES-L1
- GEO
- HEO
- SSO

Notable but less successful orbits:

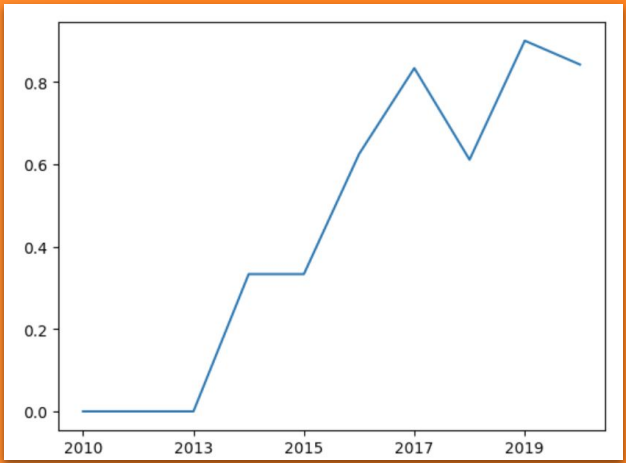
- VLEO (just above 80%)
- SSO (above 70%)
  - Source code: https://github.com/Peter-mitch1/Data-Science-Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb



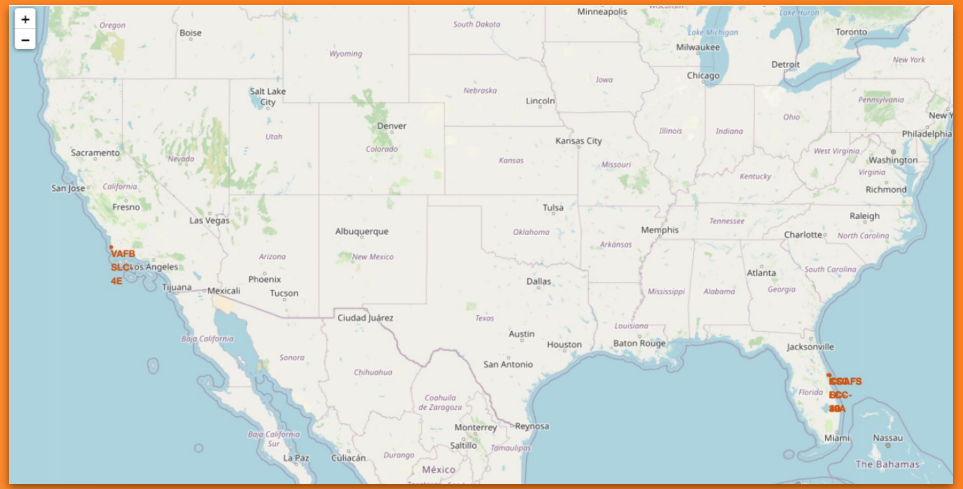
Results EDA - Launch Success Recent Yearly Trends

Success rate increased consistently from 2013-2017 dipped in 2018 and rose again in 2019 and dipped slightly in 2020

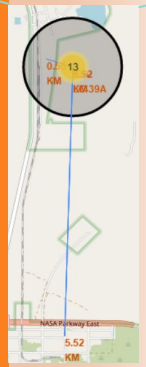
- Source code: https://github.com/Peter-mitch1/Data-Science-Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb



Results Folium - Interactive Map

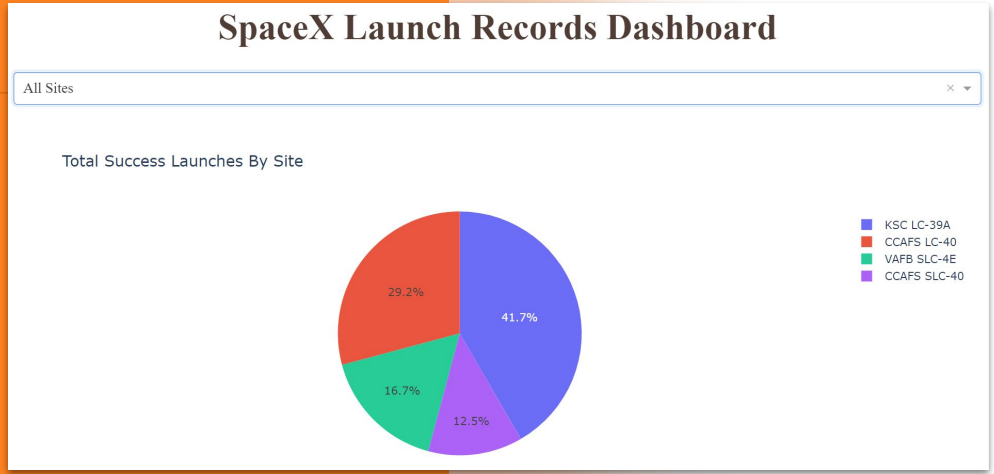


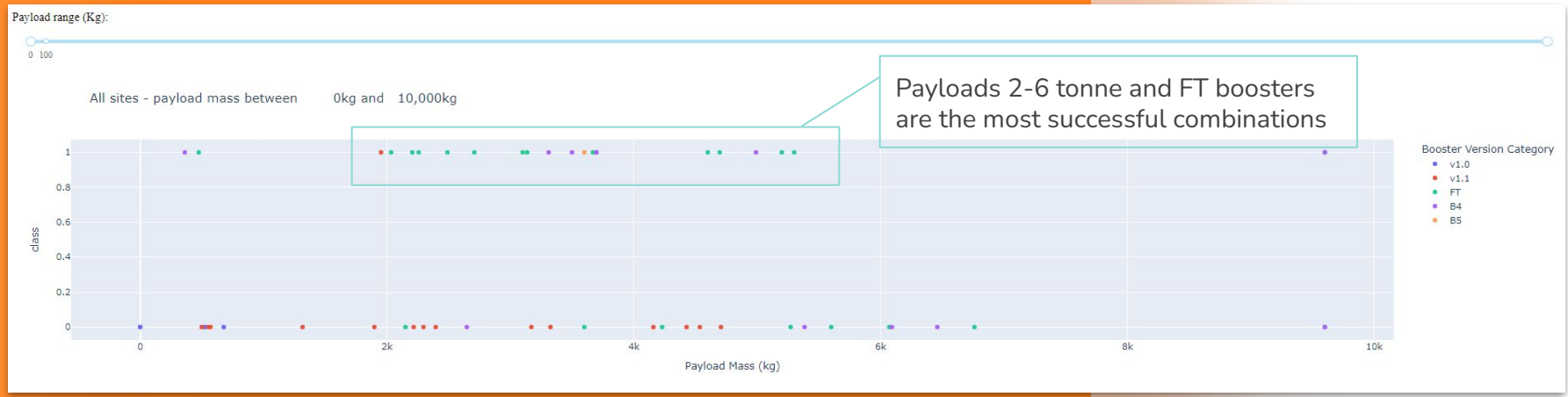
Launch site KSC LC-39A has the best logistics in regards to being close to a railroad, road and also being fairly uninhabited.



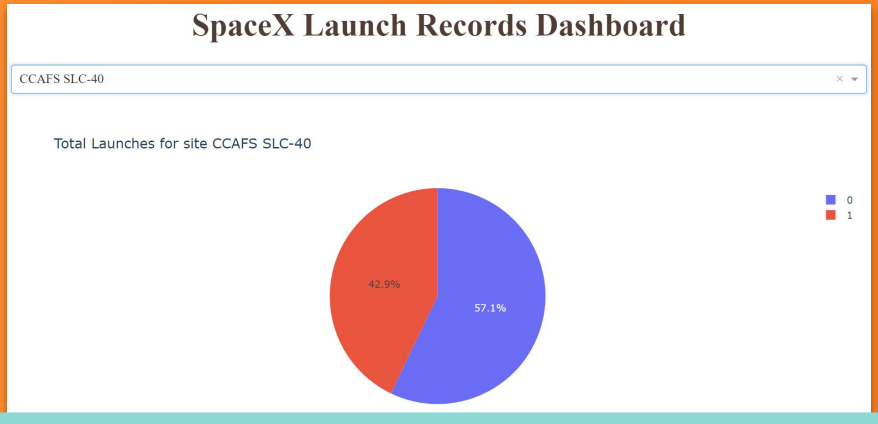
- Source code: https://github.com/Peter-mitch1/Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

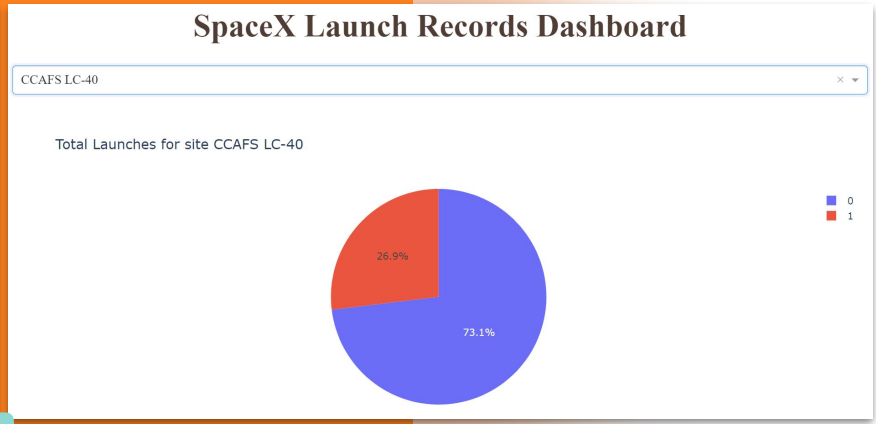
Results - Dashboard

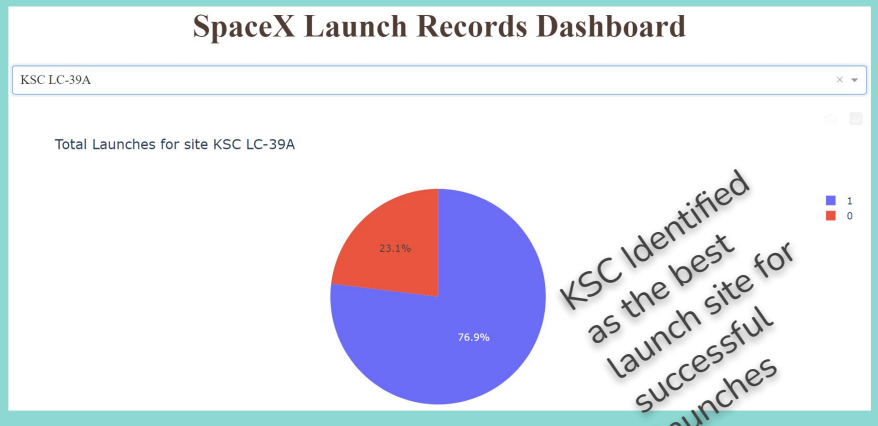


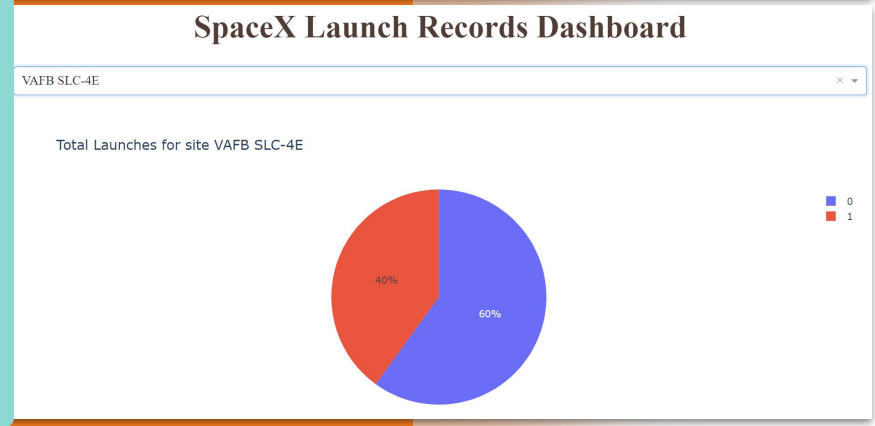


Results - Dashboard





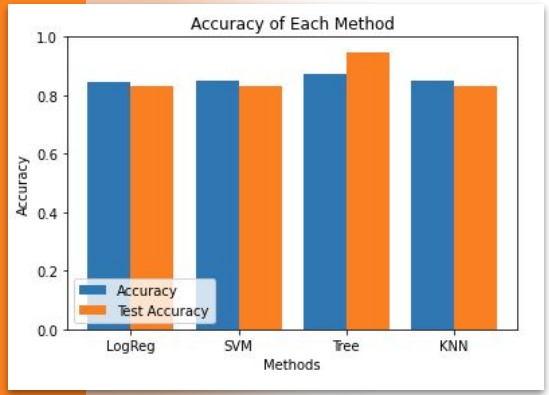




Results - Machine Learning

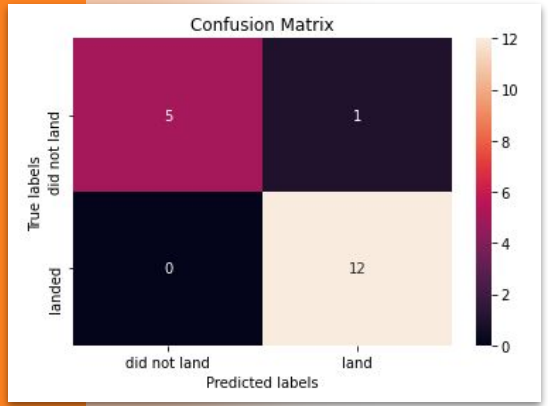
Classification Accuracy

- Four classification models tested with accuracies plotted
- The model with the highest classification accuracy was the Decision Tree Classifier with accuracy over 87%



Confusion Matrix of Decision Tree Classifier

- Confusion matrix of Decision Tree Classifier proves accuracy by large numbers of true positive and true negative compared to false positive and false negative.



Conclusion

Mission Objectives Answered:

1. Inconclusive, partly answered through finding variables that indicate successful launches such as FT boosters.

2. Answered, we can predict through machine learning with an average accuracy of around 87% the probability of successful launch.

3. Answered, The best launch site is KSC LC-39A
   (76.9% successful launches)