IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Justine Louise Remosura
9/5/2023

Outline

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

Executive Summary

- In the age of data-driven decision-making, our data science project has been instrumental in providing actionable insights and solutions to address SpaceX rocket launch program. Our project embarked on a comprehensive journey to harness the power of data, employing advanced analytics techniques to extract meaningful patterns and drive informed decisions.

- Over the course of this project, we meticulously collected, cleaned, and analyzed a diverse dataset from various sources, encompassing the scope of SpaceX rocket launch data. Through rigorous exploratory data analysis (EDA) and feature engineering, we unveiled key trends, identified critical factors impacting whether a stage 1 rocket launch phase will succeed or fail, and developed predictive models with an accuracy of 83%. These models not only enable us to predict if a rocket launch will succeed or fail but also provide a deeper understanding of the underlying dynamics.

- Our findings have far-reaching implications, including [what launch sites have the most successful stage 1 launches or what kind of rocket boosters are efficient when carrying a certain load, which can lead to improved operational efficiency, cost savings, and enhanced decision-making. We are confident that the insights derived from this data science project will be pivotal in achieving successful stage 1 rocket launch phase, and we look forward to discussing the detailed results, methodology, and next steps in this report.

Introduction

- In today's data-driven world, organizations are faced with an unprecedented influx of information. This abundance of data presents both opportunities and challenges. To harness the power of this wealth of information, businesses are increasingly turning to data science—a multidisciplinary field that combines domain expertise, statistical analysis, machine learning, and data engineering. Our data science project, titled ‘SpaceX rocket launch research’, aims to leverage these advanced techniques to solve a critical problem and extract actionable insights from a complex dataset.

- The objective of this project is to identify whether the stage 1 process of a rocket launch will succeed or fail based on historical data. As organizations strive for more informed decision-making and seek to optimize their operations, data science has emerged as a crucial tool to unlock the hidden potential within data. Through this project, we seek to address the challenges faced by SpaceX and provide solutions that are both data-driven and practical.

- In this report, we will outline our methodology, data collection process, analysis techniques, and the results of our investigation. Ultimately, our goal is to offer valuable insights and recommendations that can drive positive change and decision-making within SpaceX Technologies Corporation.

Section 1

Methodology

Methodology

- Data collection via API:
- Data Collection via WebScraping
- Perform Data Wrangling
- Perform Exploratory Data Analysis (EDA) using Visualization
- Perform Exploratory Data Analysis (EDA) using SQL
- Perform interactive visual analytics using Folium
- Perform interactive visual analytics using Plotly Dash
- Perform predictive analysis using classification models

Data Collection – API

- The data collection phase identifies and accesses relevant data sources. Our primary dataset was sourced from SpaceX, an American aerospace manufacturer and space transportation company founded by entrepreneur Elon Musk in 2002. SpaceX is headquartered in Hawthorne, California. The company's primary mission is to reduce space transportation costs and enable the colonization of Mars. The dataset encompassed rocket launch data and covered a time span between 2010 - 2020, which was deemed suitable for our analysis.
- Additionally, we supplemented our primary dataset with SpaceX rocket launch data from 2010 to 2020, such as to enrich our analysis and capture additional contextual information.
- Data was collected in CSV format and underwent numerous and rigorous quality checks to ensure consistency and integrity. Furthermore, we handled missing values by not including them in a visualization or graphs, and outliers were addressed through the removal process if deemed necessary.
- The entire data collection process adhered to ethical and legal guidelines, including obtaining necessary permissions and ensuring data privacy compliance.
- LINK: https://github.com/justineremosura/IBM_Data_Science/blob/main/Week1/SpaceX_Data_Collection_API.ipynb

Data Collection - Scraping

- In order to acquire the necessary data for our data science project, we employed web scraping techniques as a pivotal data collection strategy. Web scraping involved programmatically extracting information from various websites and web pages relevant to our research objectives. We utilized Python as our primary programming language for web scraping, along with libraries such as BeautifulSoup and Scrapy, which enabled us to navigate and parse HTML and XML structures effectively.

- Our web scraping process involved identifying the specific web sources and pages that contained the desired data, defining the structure of the target information, and writing custom scripts to extract data elements, including text, tables, and images. We also incorporated robust error-handling mechanisms to handle unexpected variations in website layouts and content.

- Ethical considerations and compliance with website terms of service were paramount throughout the web scraping process, and we ensured that we did not overload servers with excessive requests. The scraped data was subsequently cleaned, transformed, and integrated into our project's dataset, providing valuable insights for our data analysis and modeling phases.

- LINK: https://github.com/justineremosura/IBM_Data_Science/blob/main/Week1/SpaceX_WebScraping.ipynb

Data Wrangling

- Data wrangling played a crucial role in preparing our raw dataset for analysis and modeling. Our dataset, obtained from various sources, arrived in an unstructured and often messy format. The data wrangling process involved a series of systematic steps to clean, transform, and structure the data into a more suitable and usable format.
- This included handling missing values by employing techniques like imputation or removal, dealing with outliers, and addressing inconsistencies and inaccuracies in the data. We performed data type conversions to ensure consistency and compatibility among variables. Additionally, we conducted feature engineering to create new meaningful variables and extracted relevant information from text or date fields when necessary.
- Data normalization and scaling were applied to bring variables to a consistent scale for modeling purposes. This rigorous data wrangling process was essential for ensuring data quality, reducing noise, and enhancing the overall reliability of our dataset, setting the stage for robust and meaningful data analysis and modeling."
- LINK: https://github.com/justineremosura/IBM_Data_Science/blob/main/Week1/SpaceX_Data_Wrangling.ipynb

Exploratory Data Analysis with SQL

- In the initial phases of our data science project, we conducted a comprehensive Exploratory Data Analysis (EDA) to gain valuable insights into the characteristics and patterns present in our dataset. EDA served as the foundation of our data-driven decision-making process.
- Exploratory data analysis (EDA) played a pivotal role in unraveling the intricacies of our dataset. Through hands-on examination of the data, we delved into its nuances and gained a deep understanding of its characteristics. Our approach involved meticulously examining individual data points, visualizing trends, and identifying patterns that might not be apparent through automated techniques alone.
- This EDA process allowed us to detect subtle outliers and anomalies that might have been overlooked otherwise. We used domain knowledge to guide our exploration, drawing upon subject matter expertise to interpret the data in a meaningful context. By scrutinizing the dataset, we unearthed valuable insights that had a direct impact on the subsequent data preprocessing and modeling phases of our project. EDA proved indispensable in ensuring that we thoroughly explored every facet of our data, leading to a richer and more nuanced understanding of the underlying information.
- LINK: https://github.com/justineremosura/IBM_Data_Science/blob/main/Week2/EDA_SQL_SQLLite.ipynb

Exploratory Data Analysis with Data Visualization

- We began by visualizing the data through histograms, scatter plots, box plots, and other graphical representations to understand the distribution of variables, detect outliers, and identify potential relationships among features. Summary statistics, including measures of central tendency and variability, were computed to provide a high-level overview of the dataset's key attributes.

- We also conducted correlation analyses and heatmaps to explore pairwise relationships between variables. During this phase, we discovered intriguing trends, such as launch sites that have the highest number failure rate, sites that have had no successful rocket launch ever since and many more, which prompted further investigation. EDA not only enabled us to uncover critical insights but also guided our subsequent data preprocessing, feature selection, and modeling efforts, ensuring that our project was based on a solid understanding of the underlying data.

- LINK: https://github.com/justineremosura/IBM_Data_Science/blob/main/Week2/EDA_Data_Visualization.ipynb

Build an Interactive Map with Folium

- Added markers to identify the location of SpaceX launch sites on the map.
- Added color-labeled markers to see all the successful and failed launch attempts that happened on the site.
- Added a PolyLine to identify the closest coastline, railway and highway on specific sites. This will be helpful in an emergency situation like if an ambulance needs to take the shortest route on site.
- LINK: https://github.com/justineremosura/IBM_Data_Science/blob/main/Week3/Folium_LaunchSite_Location.ipynb

Build a Dashboard with Plotly Dash

- Added a dropdown to be able to select a specific Space X rocket launch site.
- Added a slider to be able to increase or decrease. This enables the user to filter the payload being carried by the boosters.
- Added a pie chart to be able to see the total number of stage 1 rocket launch attempts per site.
- Added a sub-pie chart to be able to see a percentage comparison between successful and failed rocket launch attempts.
- Added a scatter plot to be able to see which boosters successfully carried a payload on a specific orbit. Users will also be able to increase or decrease the loaded payload carried by the boosters.
- LINK: https://github.com/justineremosura/IBM_Data_Science/blob/main/Week3/Plotly_SpaceX_Launch_Records.py

Predictive Analysis (Classification)

1. Import necessary libraries.
2. Define auxiliary functions.
3. Load DataFrame.
4. Process 'Class' column to numpy array.
5. Data Standardization.
6. Split the data to train and test data.
7. Create algorithm object.
8. Load model to the object.
9. Calculate model accuracy.
10. Repeat steps 7 to 9 until we calculate the accuracy of all the chosen model algorithms.
11. Compare the accuracy results of each model algorithm to one another.
12. LINK: https://github.com/justineremosura/IBM_Data_Science/blob/main/Week4/SpaceX_ML_Prediction.ipynb

Results

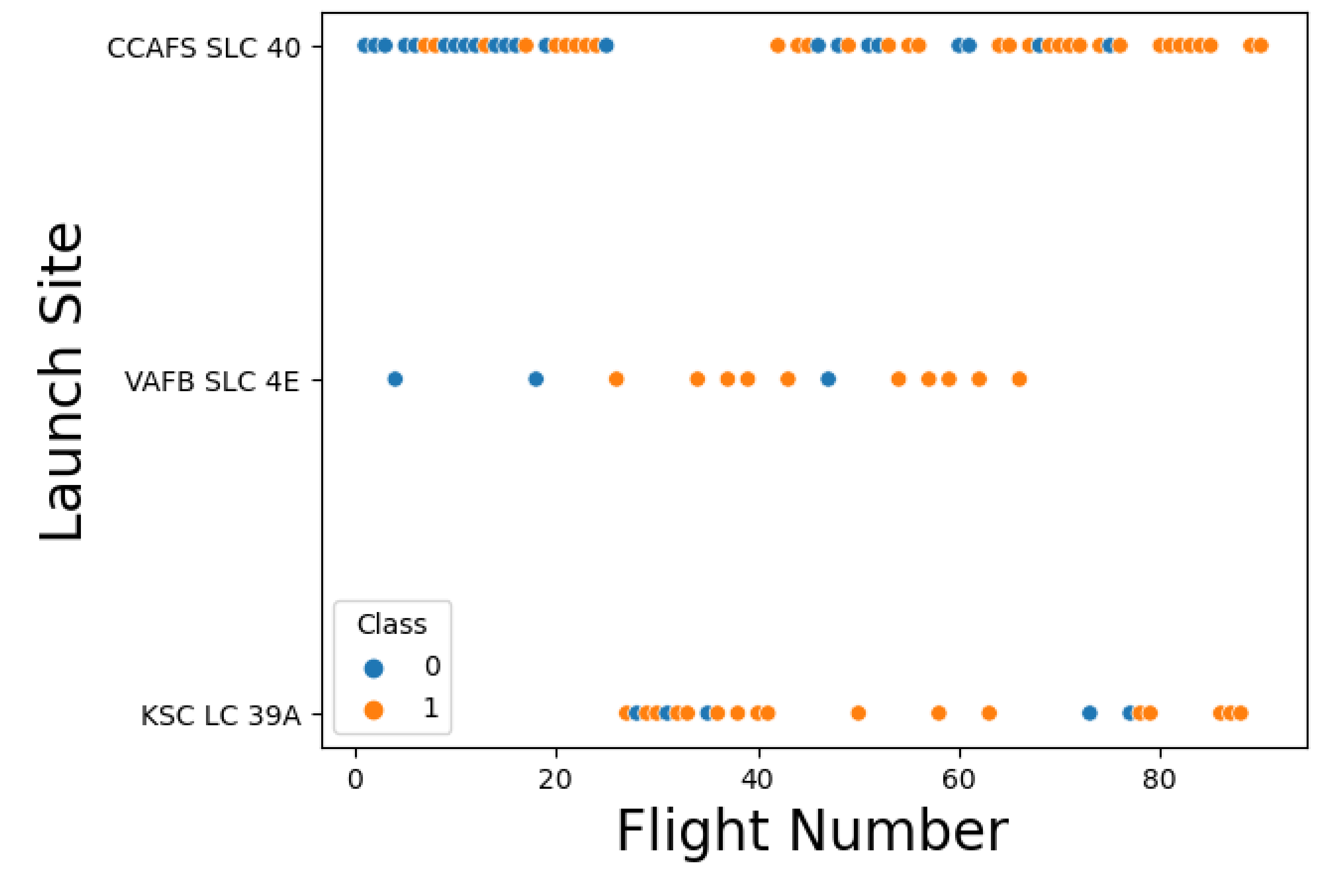
- In our data science project, predictive analysis takes center stage as we endeavor to forecast future outcomes and make proactive decisions. Leveraging a robust dataset and cutting-edge machine learning algorithms, we embarked on the journey of predictive modeling with the aim of providing valuable insights into the SpaceX rocket launch program. Our predictive models are designed to not only understand historical trends but also anticipate future events with a high degree of accuracy.
- To achieve this, we employed various machine learning techniques such as Logistic Regression, KNNs, Decision Trees and Support Vector Machine models which were selected based on their suitability for the task at hand. After an extensive feature engineering process to extract relevant information from the data, we split the dataset into training and testing sets to ensure the robustness of our models. The training data allowed our models to learn patterns, relationships, and dependencies within the data, while the testing data served as an evaluation ground for assessing model performance.
- The results of our predictive analysis are promising, with models achieving 83% accuracy. This level of accuracy empowers stakeholders to make informed decisions and implement preventive measures to mitigate potential risks or capitalize on opportunities.
- The predictive insights generated by our models can be invaluable for identifying whether the stage 1 phase of a rocket launch program will succeed or fail. As we delve deeper into the details of our analysis in subsequent sections, we will further illustrate how our predictive models can be a powerful tool in shaping the future of the SpaceX rocket launch program.

Section 2

Insights drawn from EDA

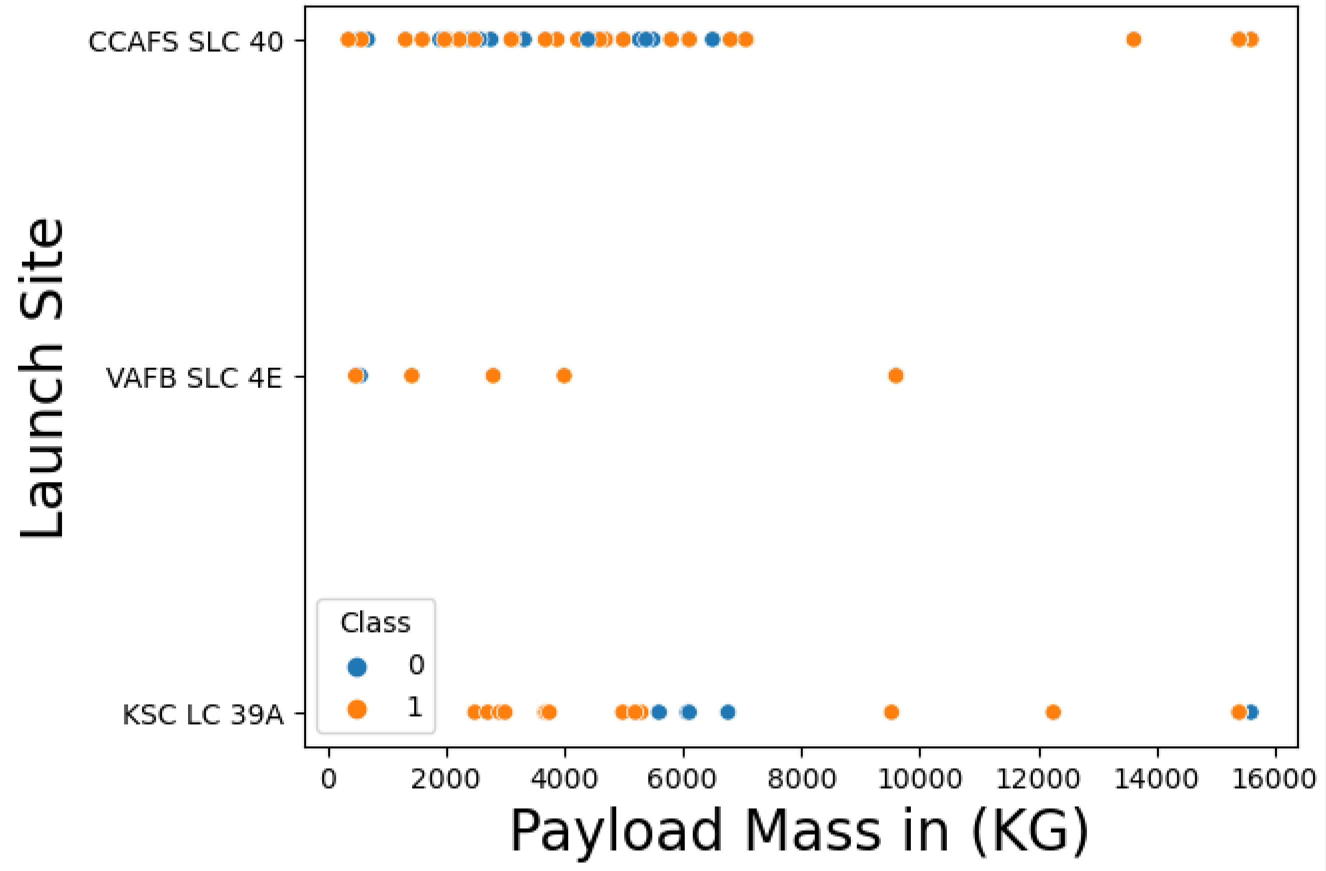
Number of flights made per sites

- SLC 40 has the highest number of launches attempted in the past.
- SLC 4E has the least amount of rocket launch attempts.
- SLC 4E only had three failed launch attempts.
- LC 39A has a more successful launch attempts record compared to SLC40 but SLC40 wins in sheer number of launch attempts.



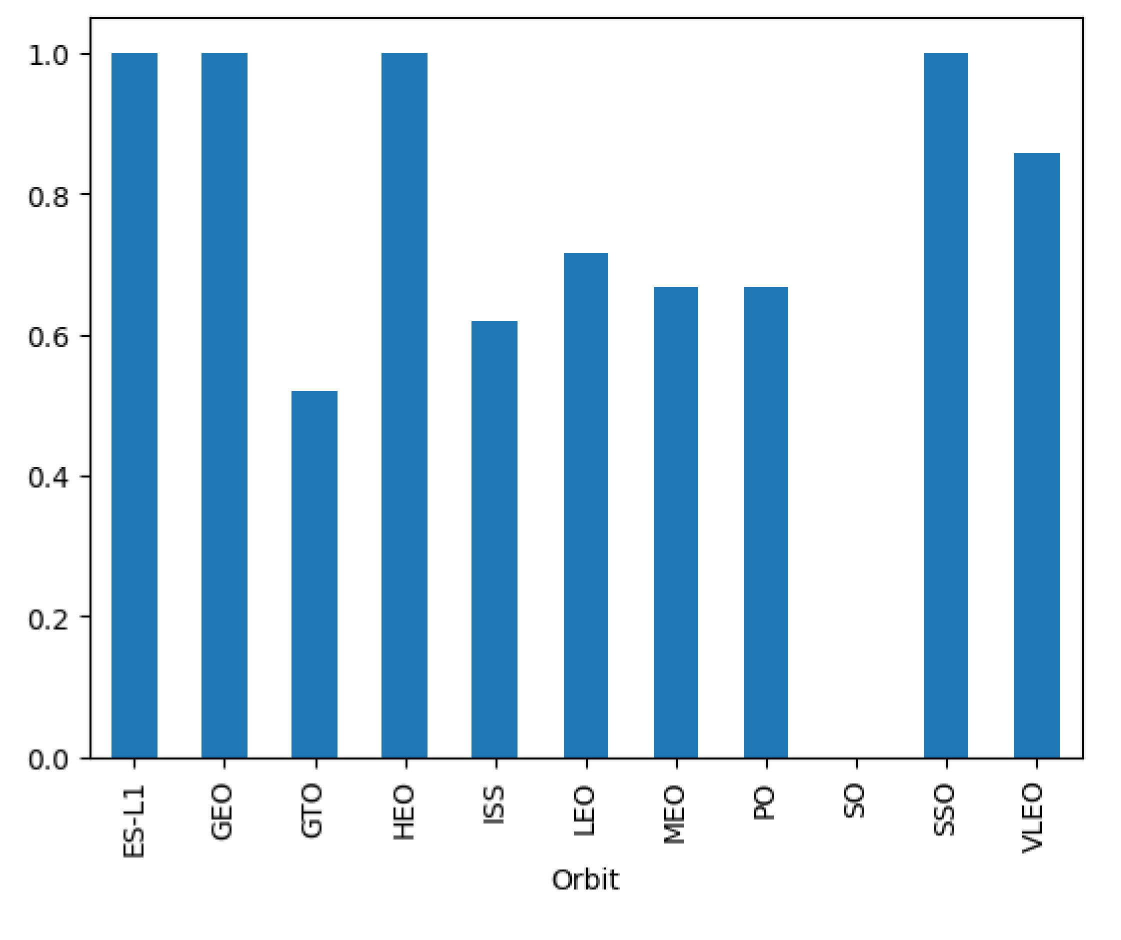
Payload to Launch Sites comparison

- SLC 40 and LC39A both maxes out at 15,800KG
- SLC 4E only carries a maximum load of 10,000KG
- Only LC39A was able to successfully launch a rocket carrying a max payload.
- SLC40 was not able to launch a rocket that carried a maximum payload.



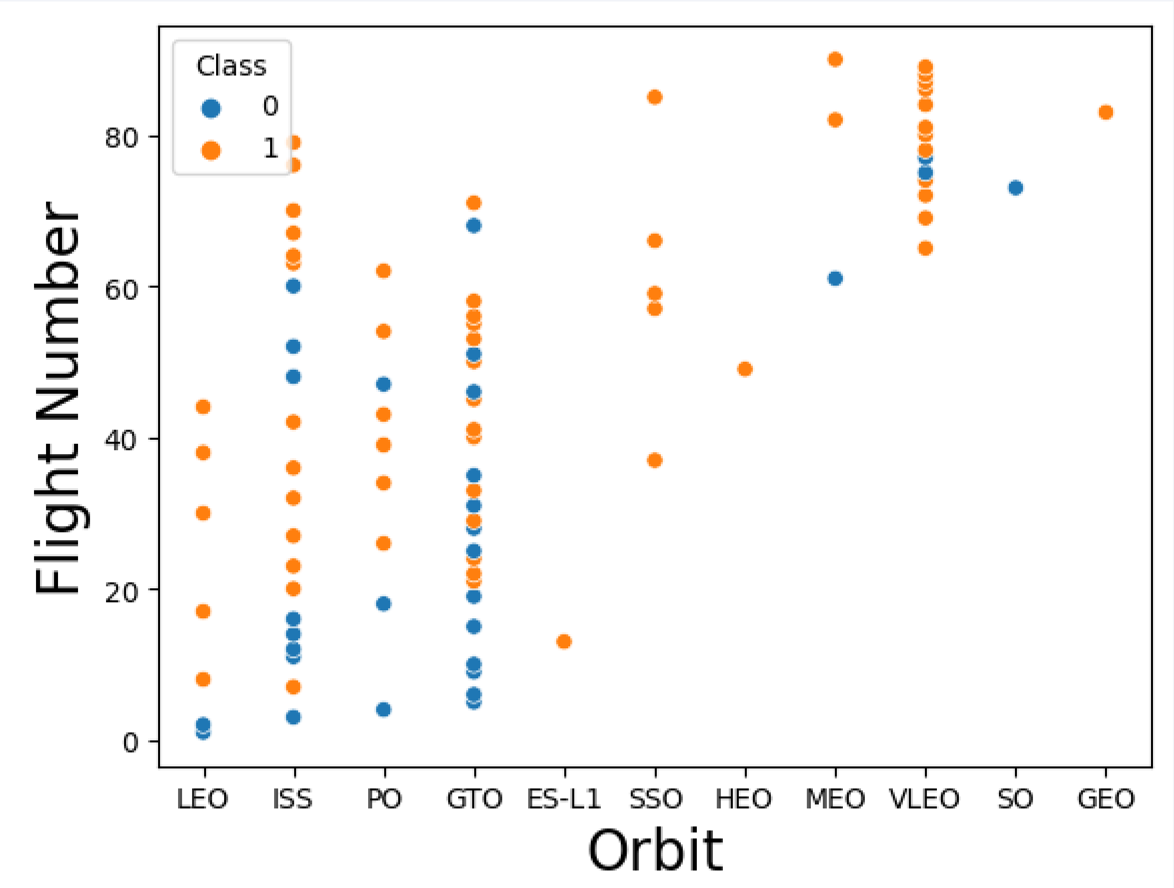
Success Rate per Orbit Type

- 1 means all rocket launch attempts succeeded. 0 means all attempts failed. Closer to 1 means most attempts succeed and closer to 0 means most attempts failed. On the following charts, we'll be able to visualize the exact launch patterns using a scatter plot.
- ESL1, GEO, HEO, SSO and VLEO are the only orbits that have a successful rocket launch.
- SO has no attempt record of launching a rocket into orbit.
- GTO, ISS, LEO, MEO and PO does not have any record of a successful rocket launch.



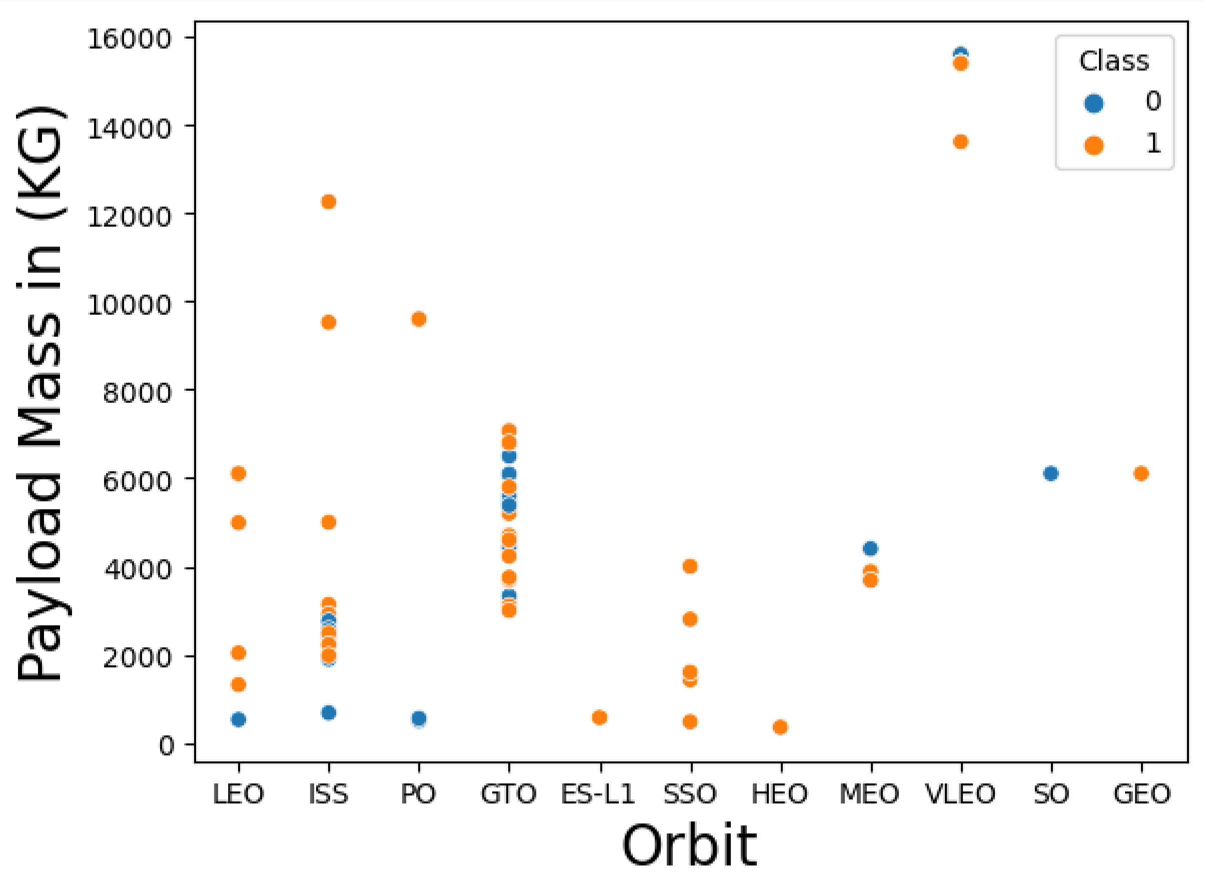
Number of flights made per orbit

- VLEO, GTO and ISS are the orbits that has the highest number of launch attempts made.
- GEO's one and only rocket launch attempt succeeded.
- ISS and GTO even though they have the most number of rocket launch attempts made also have the most number of rocket launch failures.
- SSO, GEO and HEO are the only orbits that have no record of rocket launch failure.
- Notice VLEO orbit from the previous slide. We can confirm that VLEO orbit mostly has successful rocket launch attempts made as seen from this scatter plot and the previous graph.



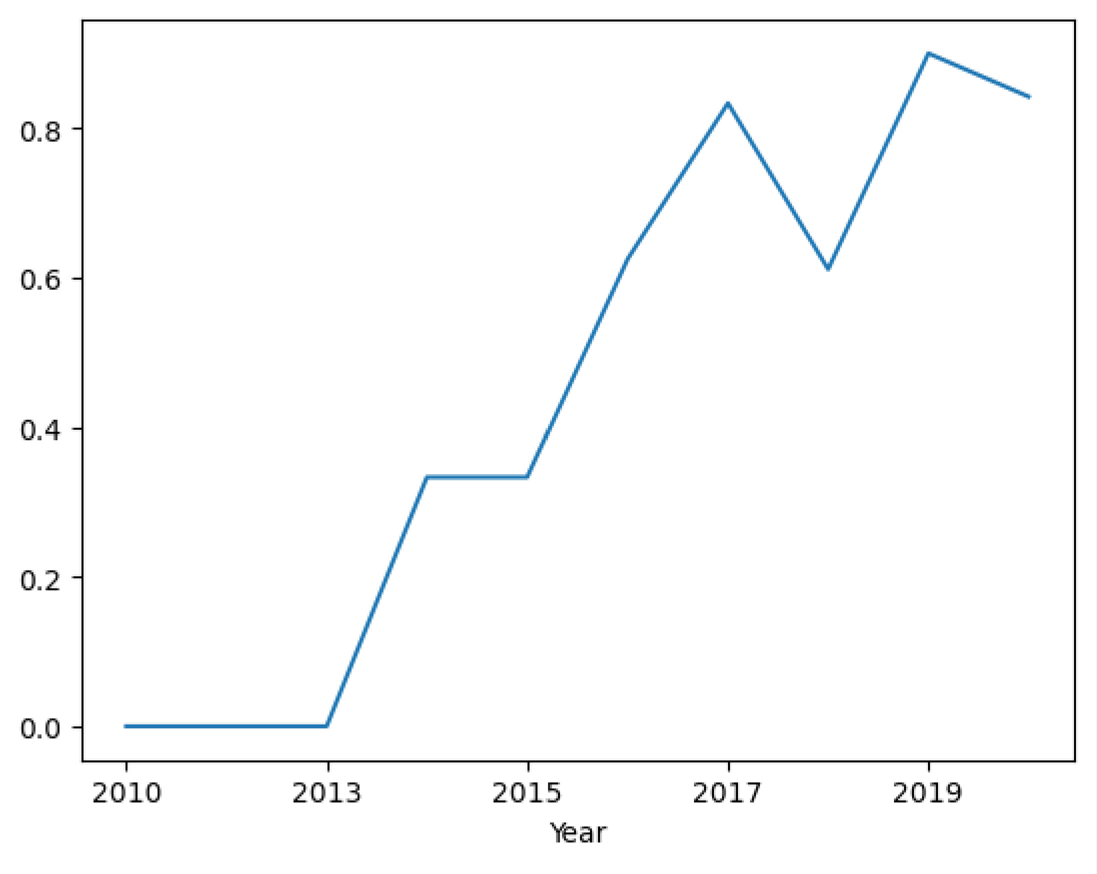
Payload to Orbit Type comparison

- VLEO orbit is the only orbit that has a successful rocket launch attempt that carries a near maximum payload.
- Rockets that go to GTO orbit carries an average payload of 5,700KG



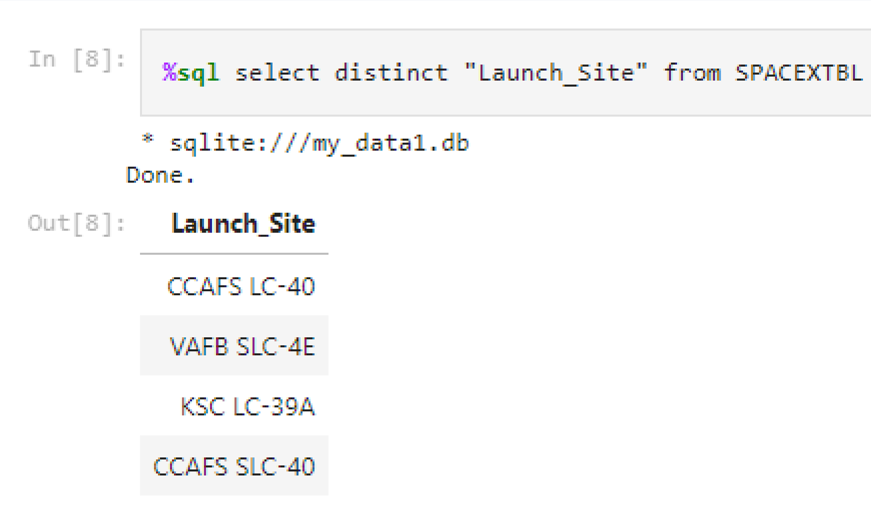
Launch Success Yearly Trend

- There’s been a steady increase in successful rocket launches since 2013.
- There was a spike in failed rocket launches around 2018.



All Launch Site Names

- This graph shows the SpaceX rocket launch sites.



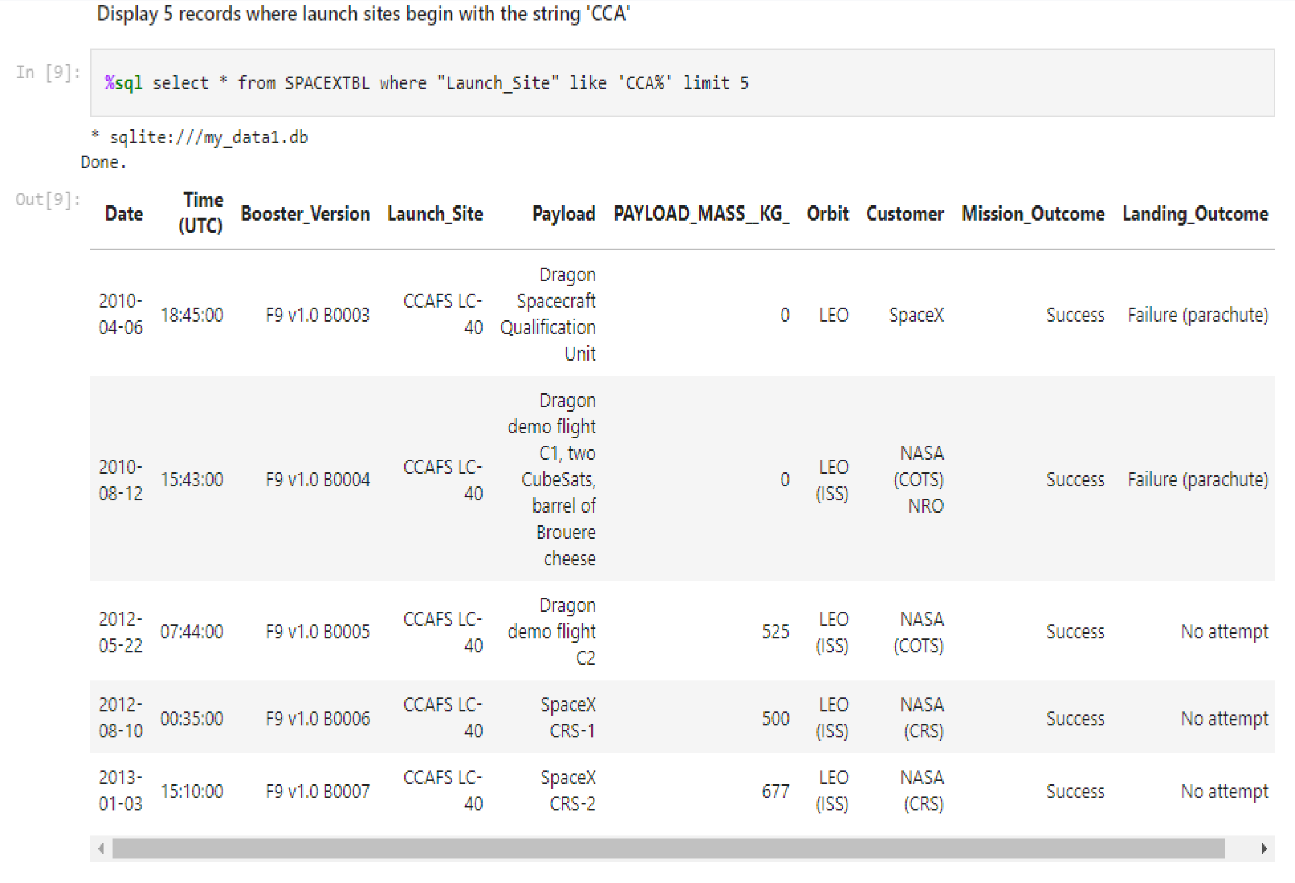
```
In [8]: %sql select distinct "Launch_Site" from SPACEXTBL

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

Out[8]:

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

Launch Site Names Begin with 'CCA'

- The image on the right shows the list of records for the launch sites that start with ‘CCA’.



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

In [9]:
```
%sql select * from SPACEXTBL where "Launch_Site" like 'CCA%' limit 5
```

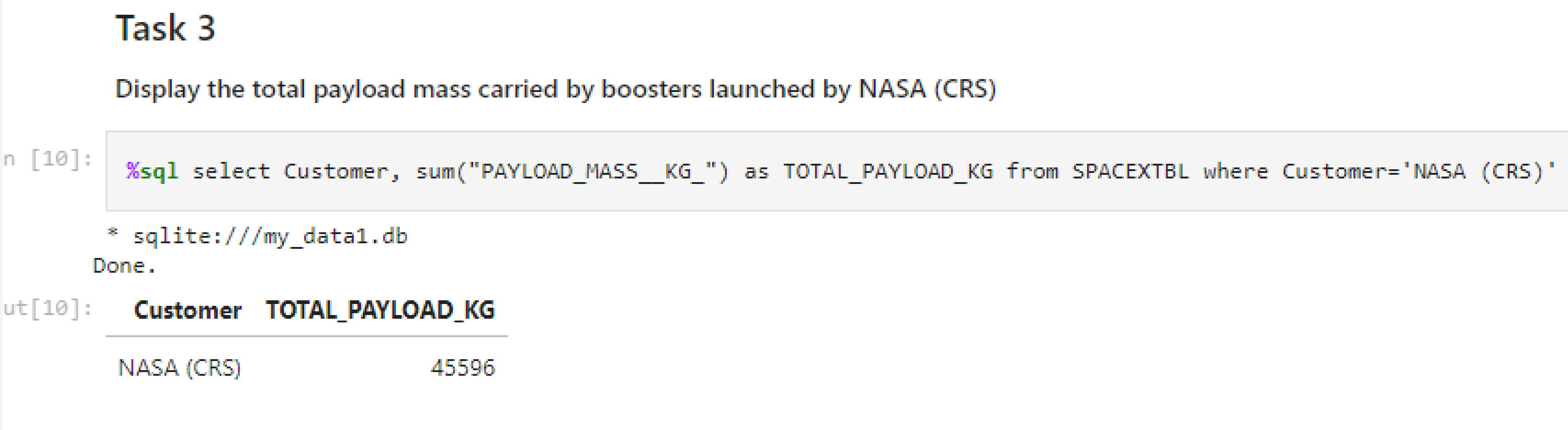
```
 * sqlite:///my_data1.db
Done.
```

Out[9]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- This image shows the total payload mass launch by NASA.



Task 3

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

n [10]:

```
%sql select Customer, sum("PAYLOAD_MASS__KG_") as TOTAL_PAYLOAD_KG from SPACEXTBL where Customer='NASA (CRS)'
```

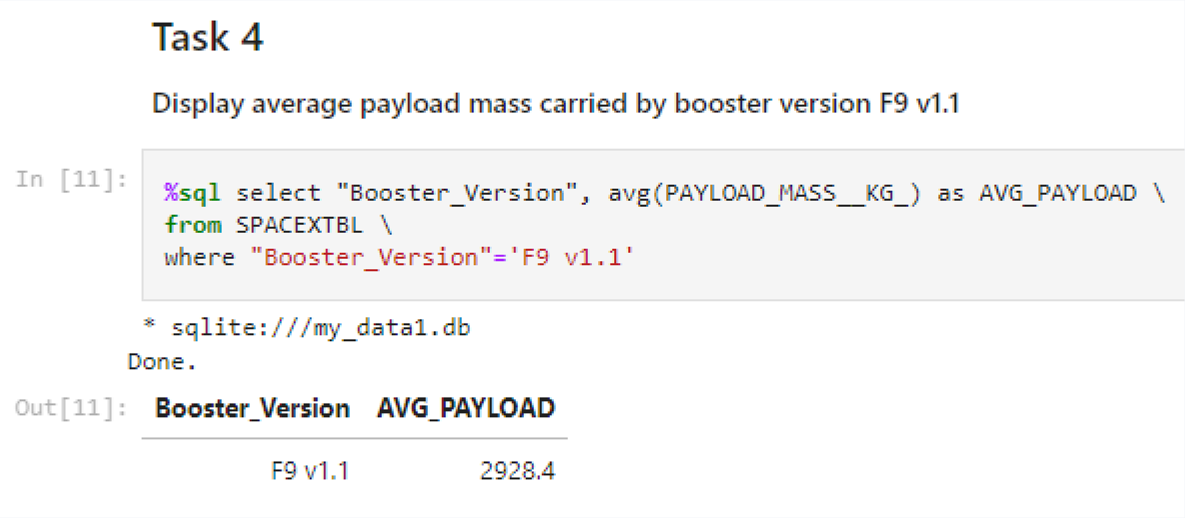
```
 * sqlite:///my_data1.db
Done.
```

ut[10]:

Customer	TOTAL_PAYLOAD_KG
NASA (CRS)	45596

Average Payload Mass by F9 v1.1

- The image shows the average payload mass carried by a Booster named ‘F9 v1.1’.



Task 4

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

```
In [11]: %sql select "Booster_Version", avg(PAYLOAD_MASS__KG_) as AVG_PAYLOAD \
         from SPACEXTBL \
         where "Booster_Version"='F9 v1.1'
```

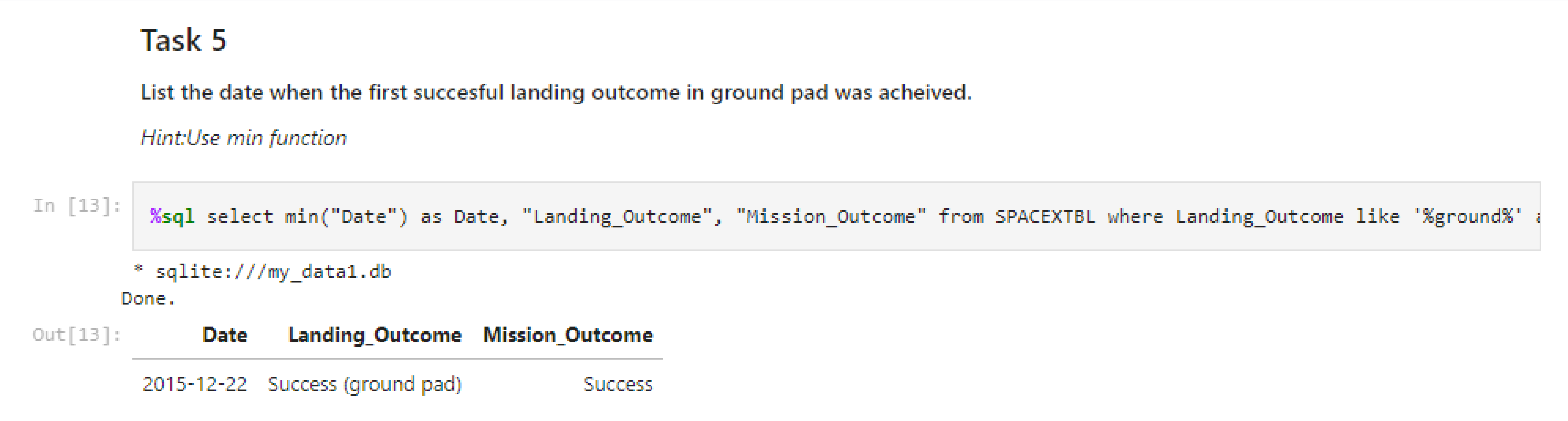
```
 * sqlite:///my_data1.db
Done.
```

Out[11]:

Booster_Version	AVG_PAYLOAD
F9 v1.1	2928.4

First Successful Ground Landing Date

- The image shows the first successful landing date achieved in a ground pad.



Task 5

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

Hint:Use min function

In [13]:

```
%sql select min("Date") as Date, "Landing_Outcome", "Mission_Outcome" from SPACEXTBL where Landing_Outcome like '%ground%' a
```

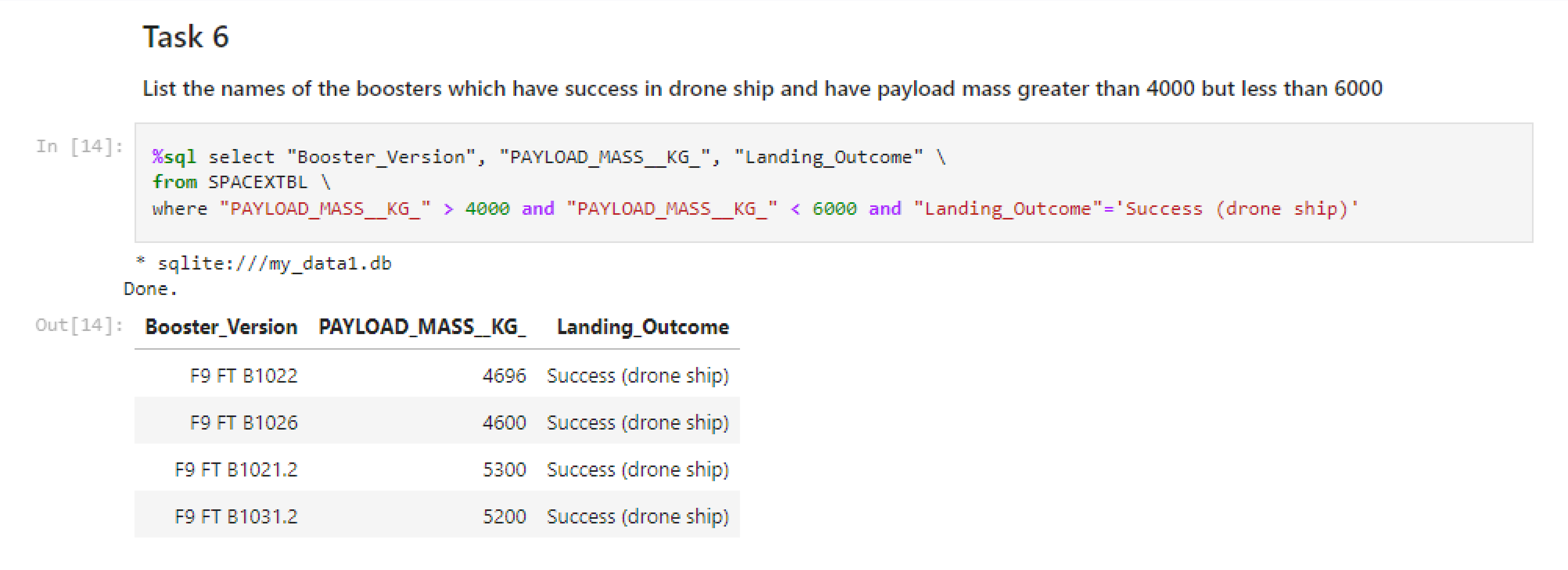
```
 * sqlite:///my_data1.db
Done.
```

Out[13]:

Date	Landing_Outcome	Mission_Outcome
2015-12-22	Success (ground pad)	Success

Successful Drone Ship Landing with Payload between 4000 and 6000

- The image shows the names of the boosters which have successfully landed on a drone ship and had payload mass greater than 4000 but less than 6000.



Task 6

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

In [14]:

```
%sql select "Booster_Version", "PAYLOAD_MASS__KG_", "Landing_Outcome" \
from SPACEXTBL \
where "PAYLOAD_MASS__KG_" > 4000 and "PAYLOAD_MASS__KG_" < 6000 and "Landing_Outcome"='Success (drone ship)'
```

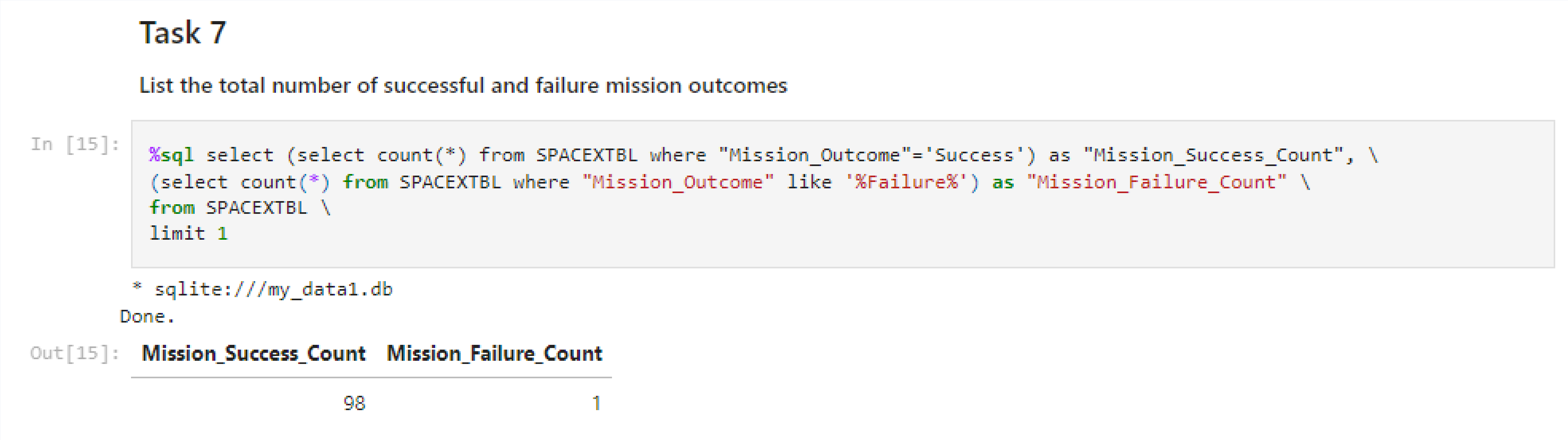
```
 * sqlite:///my_data1.db
Done.
```

Out[14]:

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

Total Number of Successful and Failure Mission Outcomes

- The image shows the total number of successful and failed mission outcomes.



Task 7

List the total number of successful and failure mission outcomes

In [15]:

```
%sql select (select count(*) from SPACEXTBL where "Mission_Outcome"='Success') as "Mission_Success_Count", \
(select count(*) from SPACEXTBL where "Mission_Outcome" like '%Failure%') as "Mission_Failure_Count" \
from SPACEXTBL \
limit 1
```

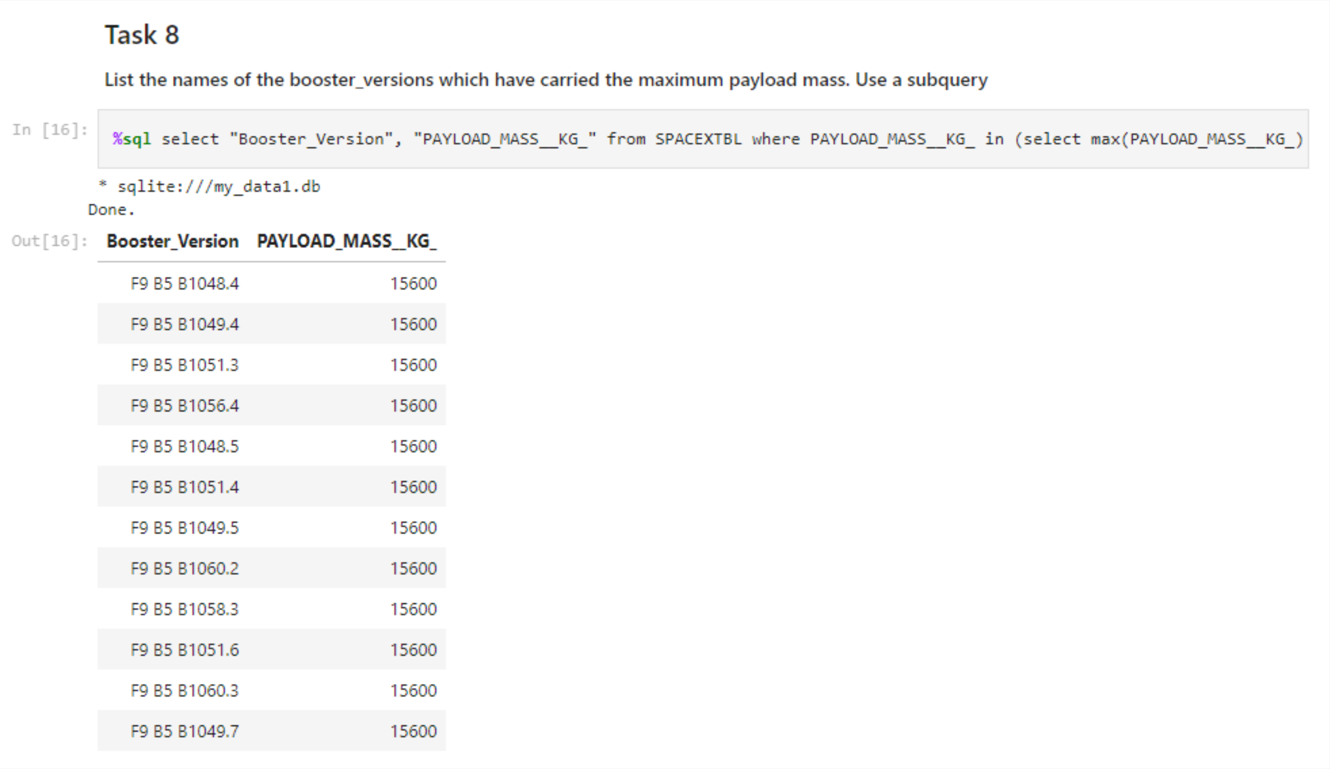
```
 * sqlite:///my_data1.db
Done.
```

Out[15]:

Mission_Success_Count	Mission_Failure_Count
98	1

Boosters Carried Maximum Payload

- The image shows the list of names of the boosters which have carried the maximum payload mass.



Task 8

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

```
In [16]: %sql select "Booster_Version", "PAYLOAD_MASS__KG_" from SPACEXTBL where PAYLOAD_MASS__KG_ in (select max(PAYLOAD_MASS__KG_)
```

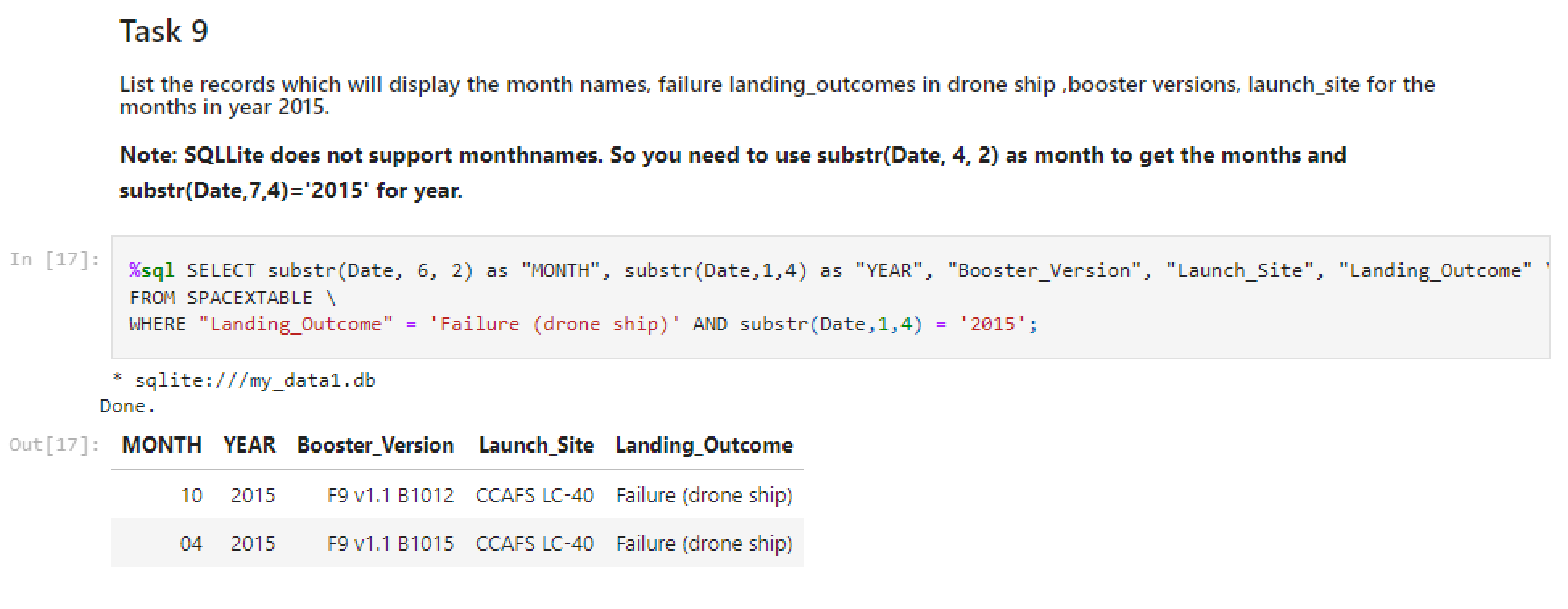
```
 * sqlite:///my_data1.db
Done.
```

Out[16]:

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

2015 Launch Records

- Image shows the list of the failed landing outcomes in drone ships, their booster versions, and launch site names for in year 2015.



Task 9

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

In [17]:

```
%sql SELECT substr(Date, 6, 2) as "MONTH", substr(Date,1,4) as "YEAR", "Booster_Version", "Launch_Site", "Landing_Outcome"
FROM SPACEXTABLE \
WHERE "Landing_Outcome" = 'Failure (drone ship)' AND substr(Date,1,4) = '2015';
```

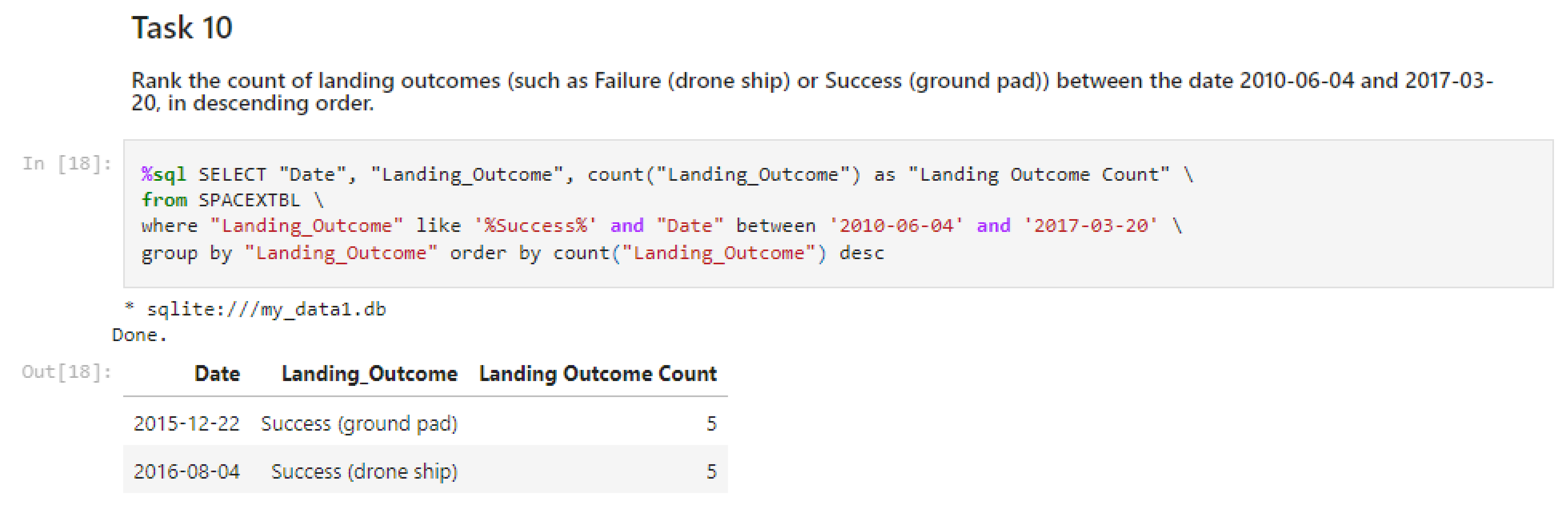
```
 * sqlite:///my_data1.db
Done.
```

Out[17]:

MONTH	YEAR	Booster_Version	Launch_Site	Landing_Outcome
10	2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- The image shows the rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the dates 2010-06-04 and 2017-03-20, in descending order.



Task 10

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

In [18]:

```
%sql SELECT "Date", "Landing_Outcome", count("Landing_Outcome") as "Landing Outcome Count" \
from SPACEXTBL \
where "Landing_Outcome" like '%Success%' and "Date" between '2010-06-04' and '2017-03-20' \
group by "Landing_Outcome" order by count("Landing_Outcome") desc
```

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

Out[18]:

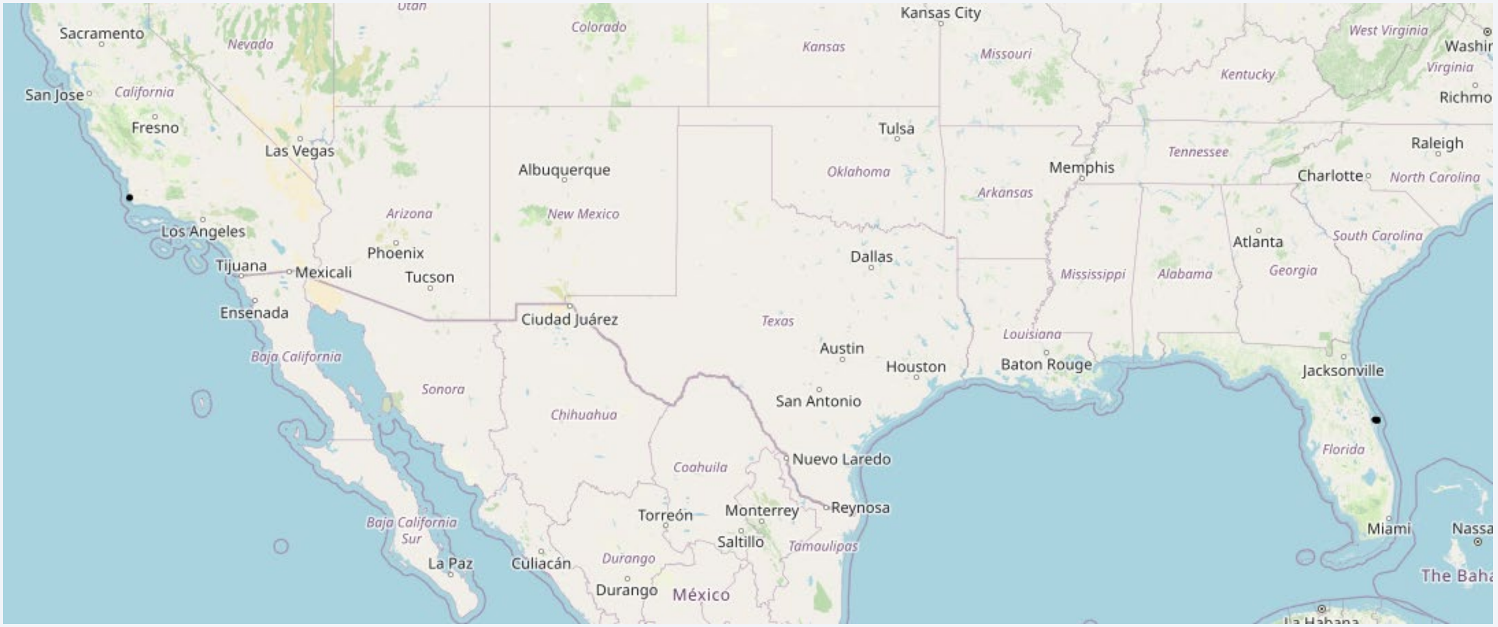
Date	Landing_Outcome	Landing Outcome Count
2015-12-22	Success (ground pad)	5
2016-08-04	Success (drone ship)	5

Section 3

Launch Sites Proximities Analysis

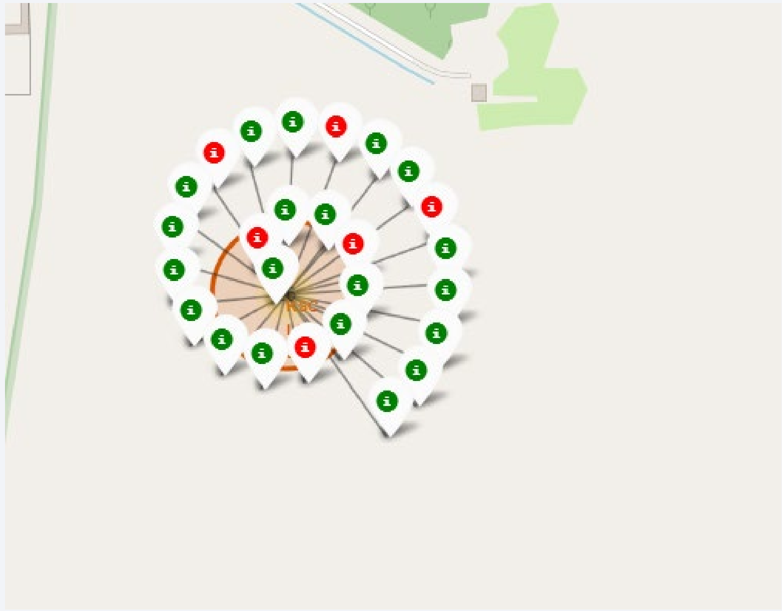
SpaceX launch site locations

- The two black dots on the map are the SpaceX launch site locations.
- It is important to apply appropriate zoom levels on the map to be able to see all the tagged locations.



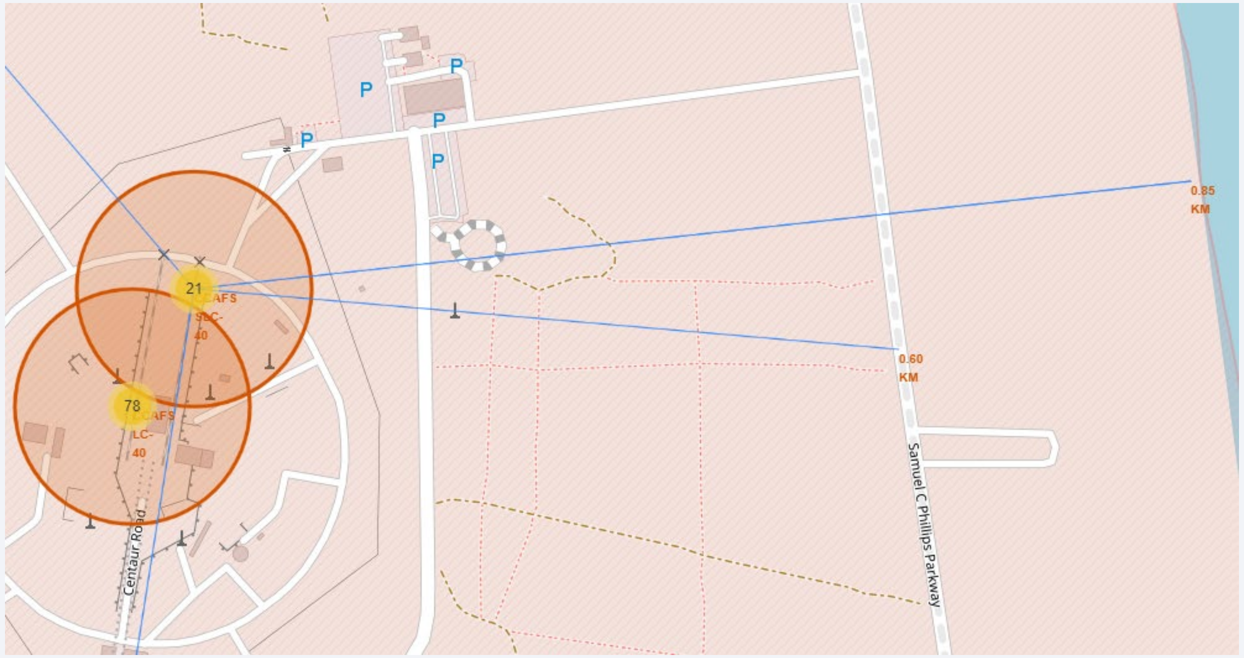
Visual representation for successful and failed launches

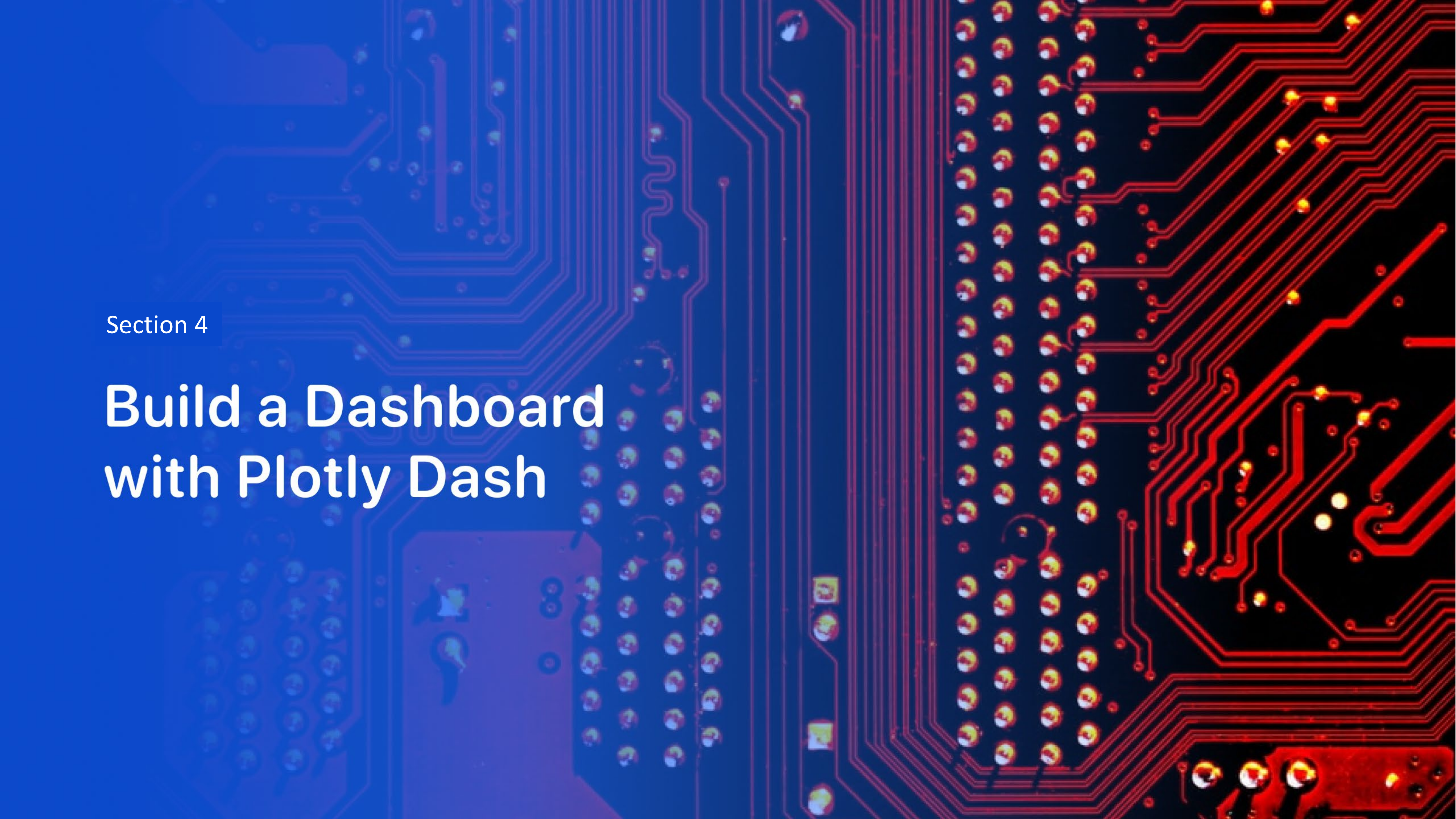
- LC39A will be used as an example. In this image, we can see the successful and failed rocket launch that happened on the LC39A launch site.
- Green represents a successful rocket launch and red means a failed launch.



Nearest coastline, city, railway & highway on a launch site

- We will be using the SLC40 Launch Site as an example. In this image, we can identify which coastline and what highway is the nearest to SLC40 Launch Site. This will be useful for emergency situations like if an ambulance or emergency personnel needs to find the nearest route on site.



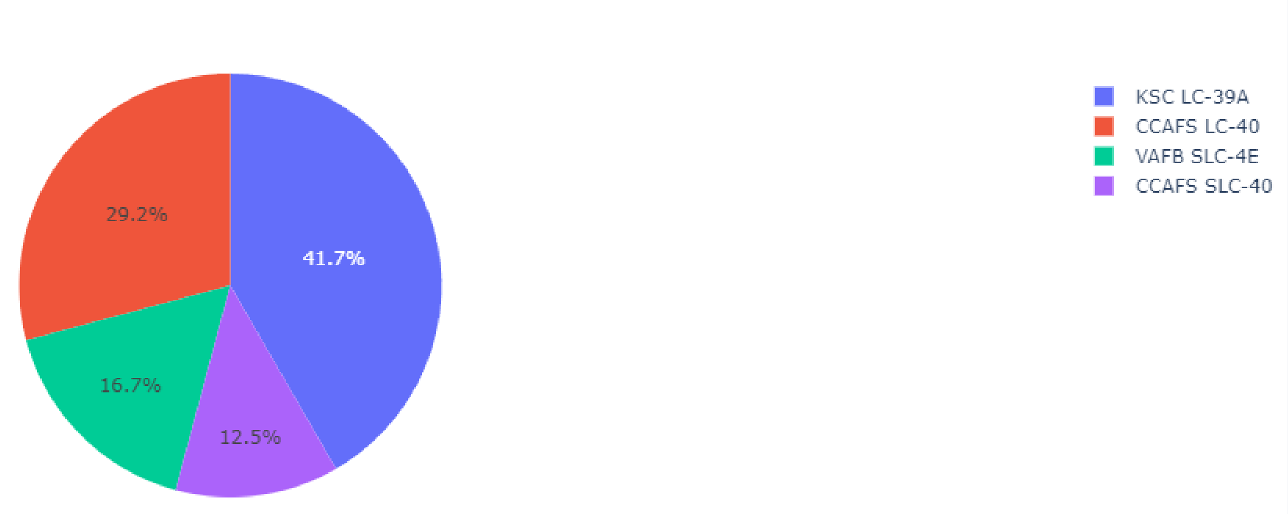


Section 4

Build a Dashboard with Plotly Dash

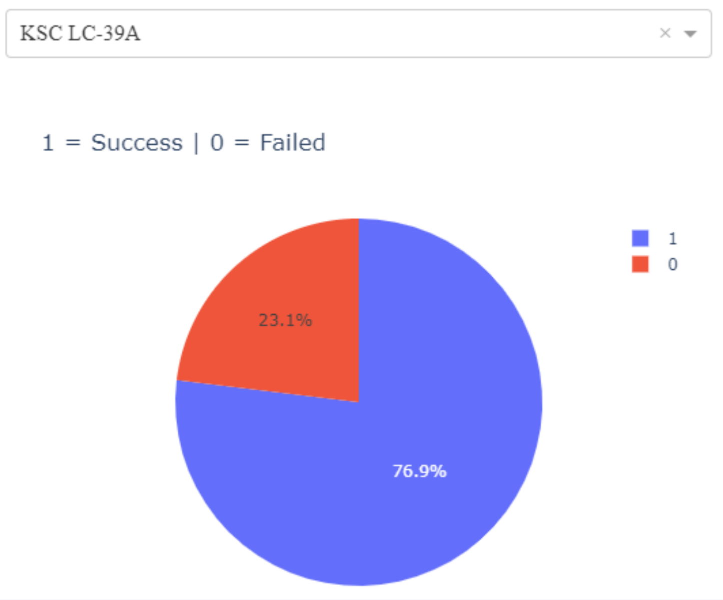
Total launch attempts per launch sites

- LC39A has the highest amount of rocket launch attempts followed by LC40.
- This graph does not identify whether a rocket launch attempt is successful or not.
- The site that has the least amount of rocket launch attempts is SLC40 followed by SLC4E.
- This pie graph only indicates the total number of rocket launch attempts per launch site.



Pie Chart for LC39A Site

- In this image, we check the percentage of success rate compared to failure rate of the launch site LC39A.
- The image on the previous slide only shows the total launch attempts per site. Selecting a specific site will instead show a pie graph comparing the successful and failed launch attempts at the selected site.



KSC LC-39A

1 = Success | 0 = Failed

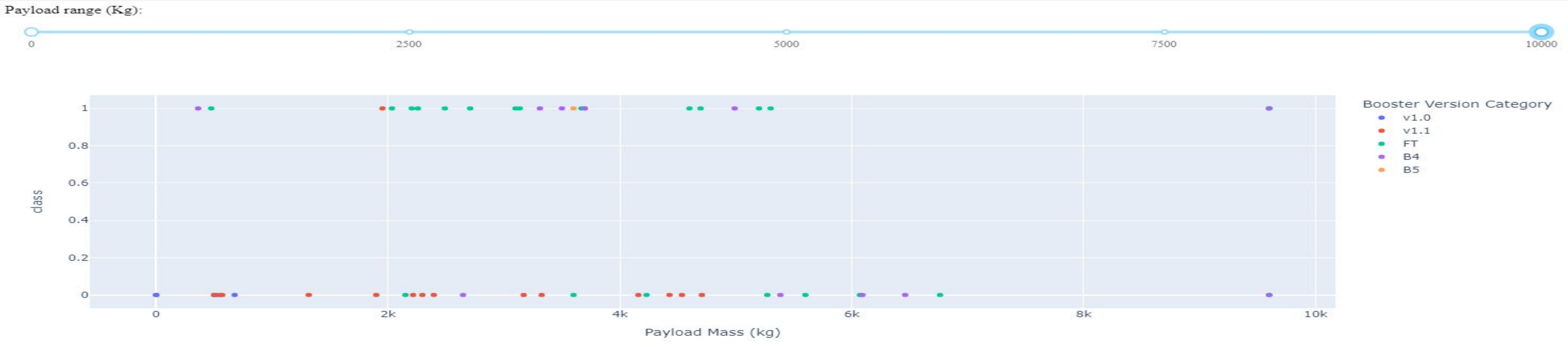
Launch Outcome per Payload

- This scatter plot applies to all launch sites. It shows which boosters have the highest success and failure rate according to the payload it's carrying.

- In the graph, we can see that the FT Booster performs the best when carrying a payload at around 2,000KG – 6,000KG. We can say that it's a booster tailored for carrying middleweight payloads.

- The B4 booster is the most capable one as it is the only booster capable of carrying payloads at around 9,000 to 10,000KG mark.

- The v1.1 booster has the highest number of failure rates.

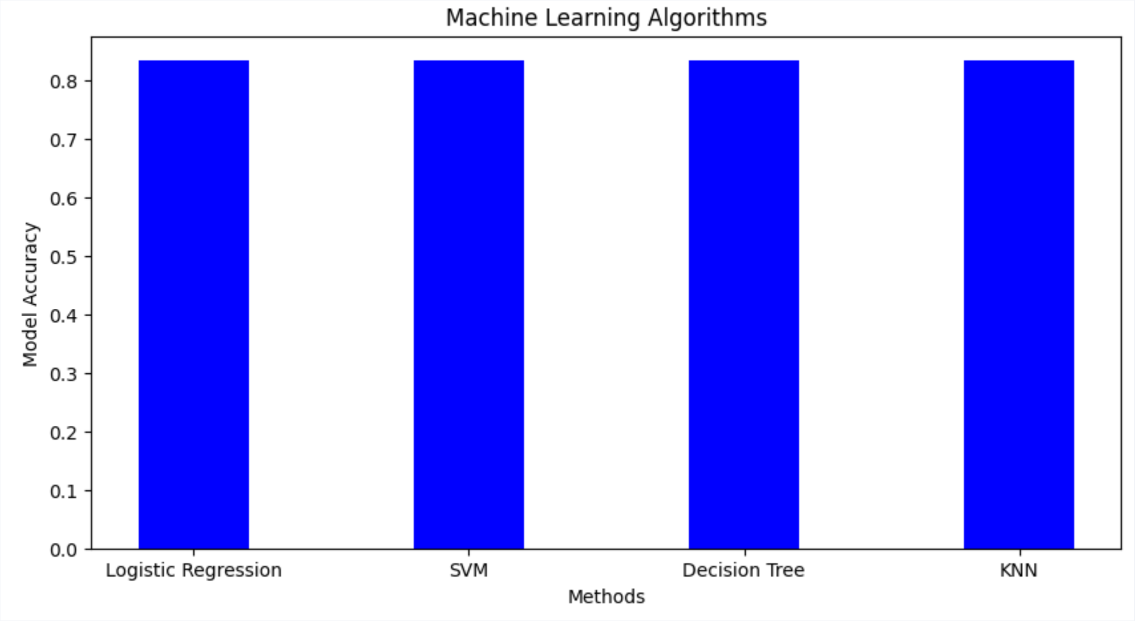


Section 5

Predictive Analysis (Classification)

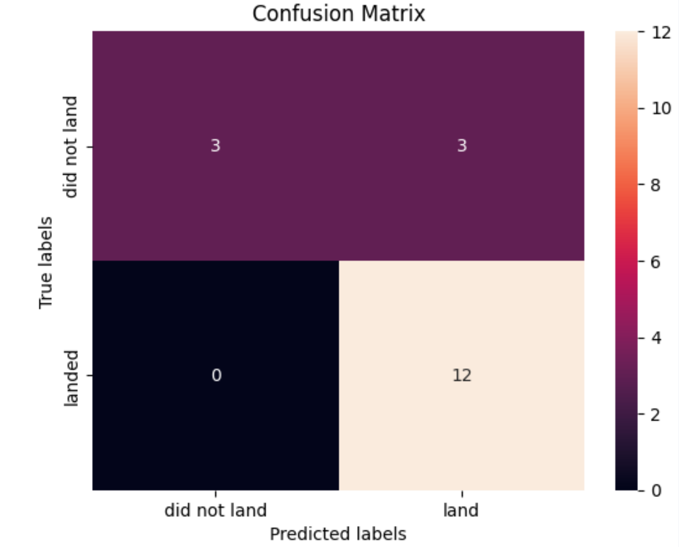
Classification Accuracy

- All four machine learning algorithms have the same accuracy.
- This determines that there is an 83% chance of success for stage 1 rocket launch.



Confusion Matrix

- All four machine learning algorithms have the same confusion matrix.
- According to the confusion matrix outputted by the algorithms, we can say that the model is accurate as it was able to predict True Positive in a fairly accurate manner.
- Top Left = True Negative, Top Right = False Positive, Lower Left = False Negative, Lower Right = True Positive
- FP and FN are indicators of error in prediction. A good model is one which has high TP and TN rates, while low FP and FN rates.



Conclusions

- All of the four models used have 83% accuracy.
- All four model algorithms have been put in a confusion matrix and are able to determine True Positives accurately. It means it does not delineate too far if a stage 1 rocket launch phase will succeed or fail in the process.
- False Positives and False Negatives only have 0 and 3 hits which means the models do not falsely tag events.

Appendix

- https://github.com/justineremosura/IBM_Data_Science/tree/main

Thank you!