IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Shuhao Zhang>

<30/09/2022>

GitHub link: https://github.com/icepoloz/Python.git

Outline

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

Executive Summary

- Collect data from public SpaceX API and Wikipedia page.
- Creat labels column ‘class’ classifying successful landing.
- Explore date using SQL, visualization, maps, and dashboards.
- Gather related columns to be used as features.
- Change all categorical variables to binary using hot encoding
- Process data and use GridSearchCV to look for best parameter for machine learning
- Visualize the data
- **Use 4 machine learning models**
- **All models can predict successful landings**

Introduction

Project background:

SpaceY needs to compete with best company, SpaceX, in this industry

Problems you:

Train a machine learning model to predict successful Stage 1 recovery

Section 1

Methodology

Methodology

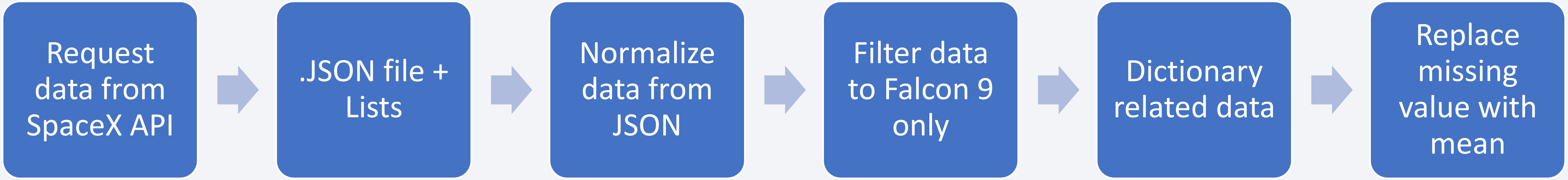
Executive Summary

- Data collection methodology:
  - Collect data from SpaceX public API and SpaceX Wikipedia page
- Perform data wrangling
  - Classifying landings as successful and failed.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Tuned models using GridSearchCV

Data Collection

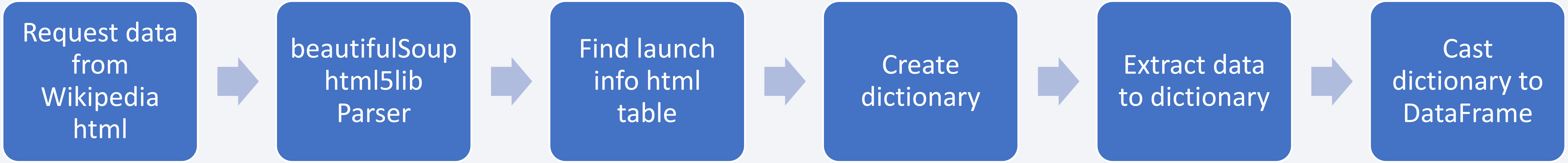
- Data collection process includes API requests and web scraping data.

Data Collection – SpaceX API


Request data from SpaceX API
.JSON file + Lists
Normalize data from JSON
Filter data to Falcon 9 only
Dictionary related data
Replace missing value with mean

https://github.com/icepoloz/Python/blob/a91300adf3e7c2d4621cd114e4bddbacf22da39d/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping



Request data from Wikipedia html → beautifulSoup html5lib Parser → Find launch info html table → Create dictionary → Extract data to dictionary → Cast dictionary to DataFrame

https://github.com/icepoloz/Python/blob/a91300adf3e7c2d4621cd114e4bddbacf22da39d/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Exploratory Data Analysis performed on variables Flight Number, Payload Mass, Launch Site, Orbit, Class and Year.
- Plots Used: Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit vs. Success Rate, Flight Number vs. Orbit, Payload vs Orbit, and Success Yearly Trend
- Scatter plots, line charts, and bar plots were used to compare relationships between variables to
- Decide if a relationship exists so that they could be used in training the machine learning model
- https://github.com/icepoloz/Python/blob/a91300adf3e7c2d4621cd114e4bddbacf22da39d/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- Loaded data set into IBM DB2 Database.
- Queried using SQL Python integration.
- Queries were made to get a better understanding of the dataset.
- Queried information about launch site names, mission outcomes, various pay load sizes of customers and booster versions, and landing outcomes
- https://github.com/icepoloz/Python/blob/a91300adf3e7c2d4621cd114e4bddbacf22da39d/sql%20notebook.ipynb

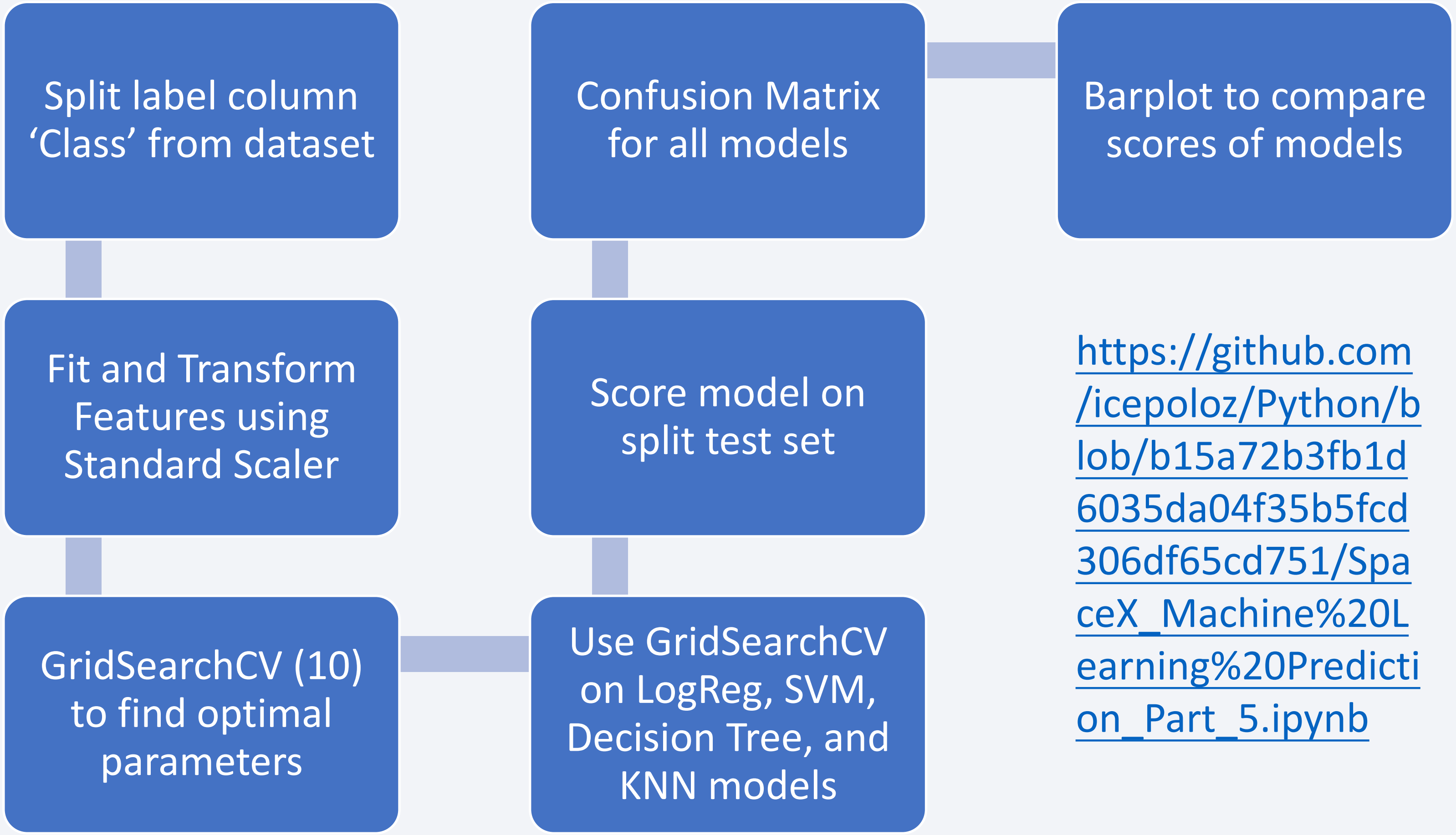
Build an Interactive Map with Folium

- Folium maps mark Launch Sites, successful and unsuccessful landings, and a proximity example to key locations: Railway, Highway, Coast, and City.
- This allows us to understand why launch sites may be located where they are. Also visualizes successful landings relative to location.
- https://github.com/icepoloz/Python/blob/b15a72b3fb1d6035da04f35b5fcd306df65cd751/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Dashboard includes a pie chart and a scatter plot.
- Pie chart can be selected to show distribution of successful landings across all launch sites and can be selected to show individual launch site success rates.
- Scatter plot takes two inputs: All sites or individual site and payload mass on a slider between 0 and 10000 kg.
- The pie chart is used to visualize launch site success rate.
- The scatter plot can help us see how success varies across launch sites, payload mass, and
- booster version category.
- https://github.com/icepoloz/Python/blob/b15a72b3fb1d6035da04f35b5fcd306df65cd751/jupyter-labs-eda-dataviz.ipynb

Predictive Analysis (Classification)



Split label column 'Class' from dataset

Fit and Transform Features using Standard Scaler

GridSearchCV (10) to find optimal parameters

Use GridSearchCV on LogReg, SVM, Decision Tree, and KNN models

Score model on split test set

Confusion Matrix for all models

Barplot to compare scores of models

https://github.com/icepoloz/Python/blob/b15a72b3fb1d6035da04f35b5fcd306df65cd751/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

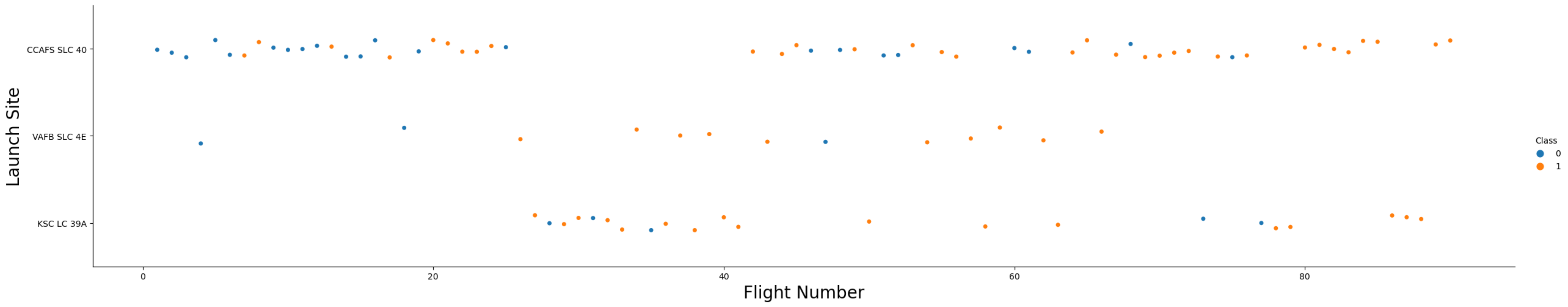
Results

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

Section 2

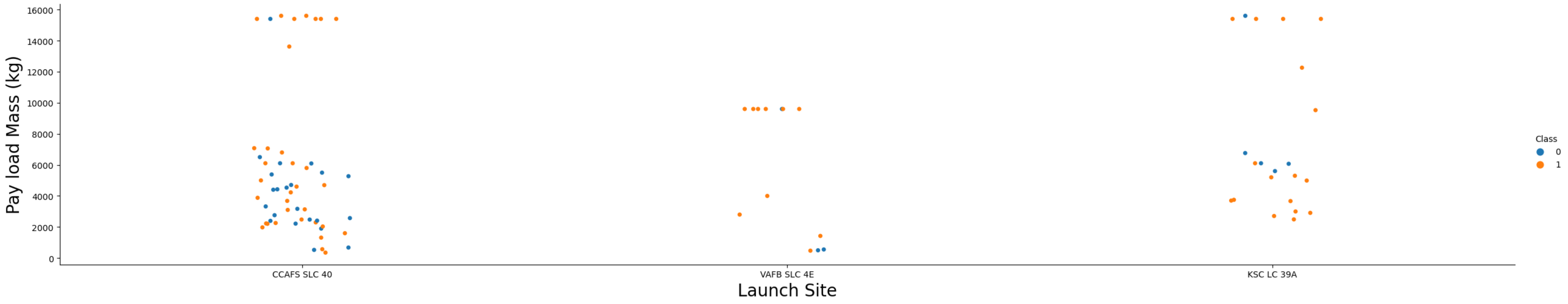
Insights drawn from EDA

Flight Number vs. Launch Site



- Orange means successful launch
- Blue means unsuccessful launch
- There is an increase in success rate over time (indicated in Flight Number). Likely a big breakthrough around flight 20 which significantly increased success rate. CCAFS appears to be the main launch site as it has the most volume.

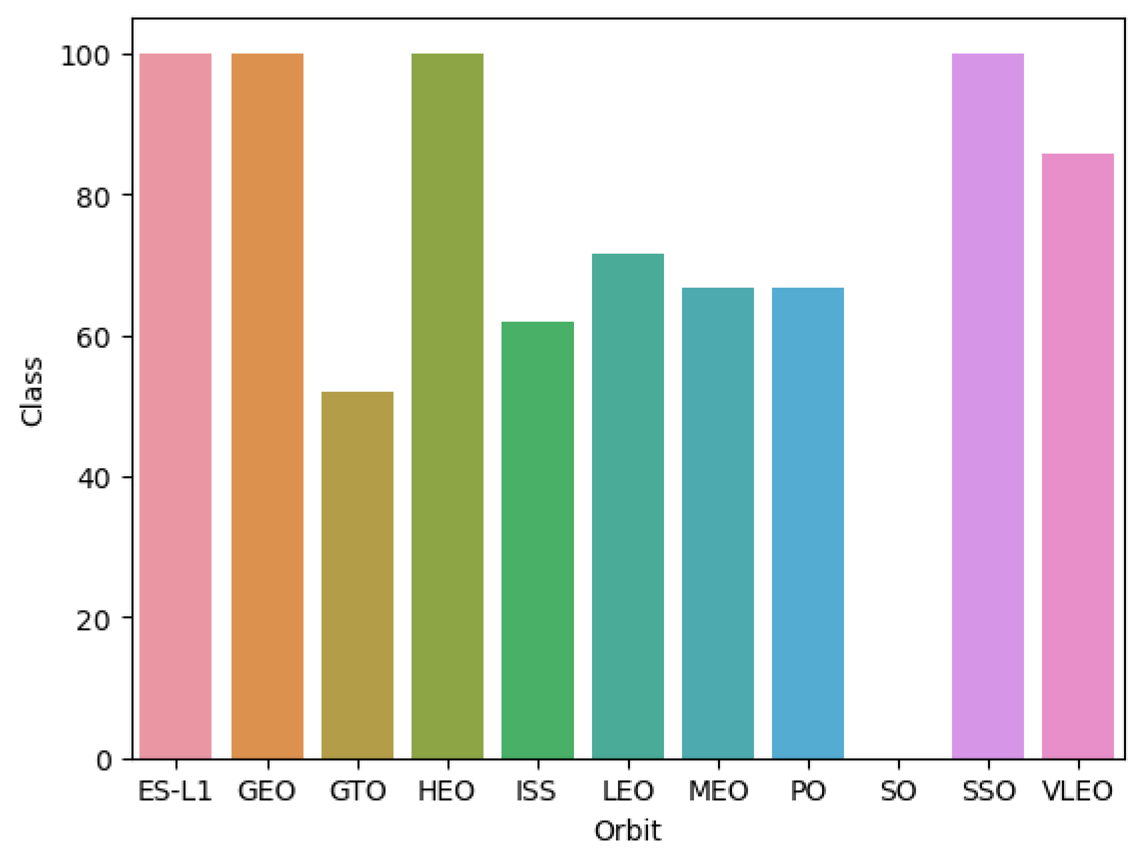
Payload vs. Launch Site



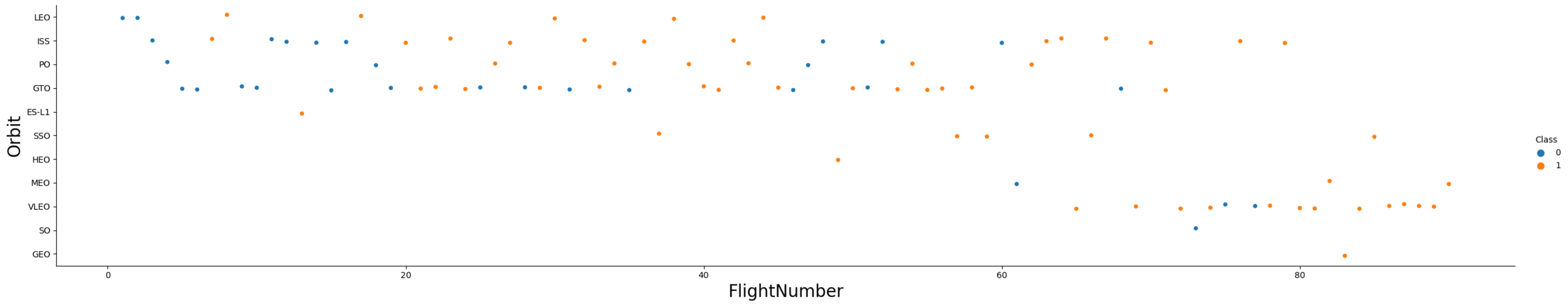
- Orange means successful launch
- Blue means unsuccessful launch
- For the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type

- ES-L1 (1), GEO (1), HEO (1) have 100% success rate (sample sizes in parenthesis) SSO (5) has 100% success rate
- VLEO (14) has decent success rate and attempts
- SO (1) has 0% success rate
- GTO (27) has the around 50% success rate but largest sample

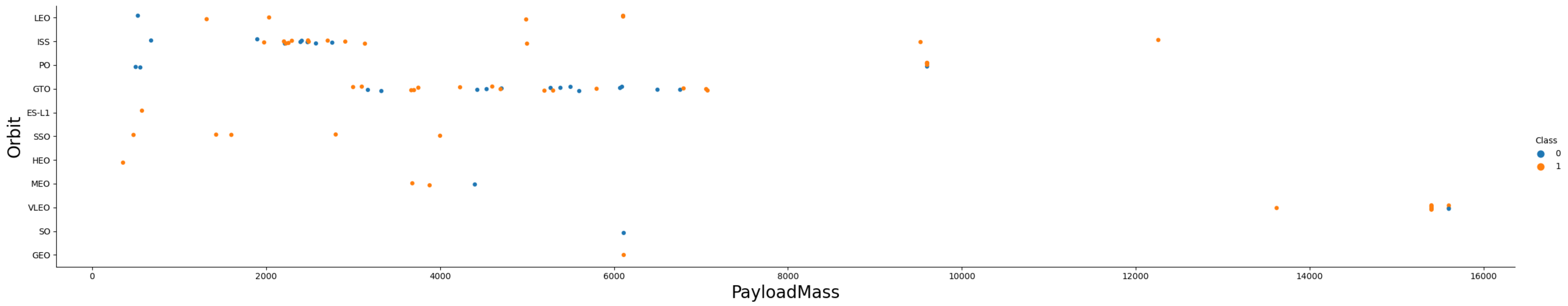


Flight Number vs. Orbit Type



- Orange means successful launch
- Blue means unsuccessful launch
- 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

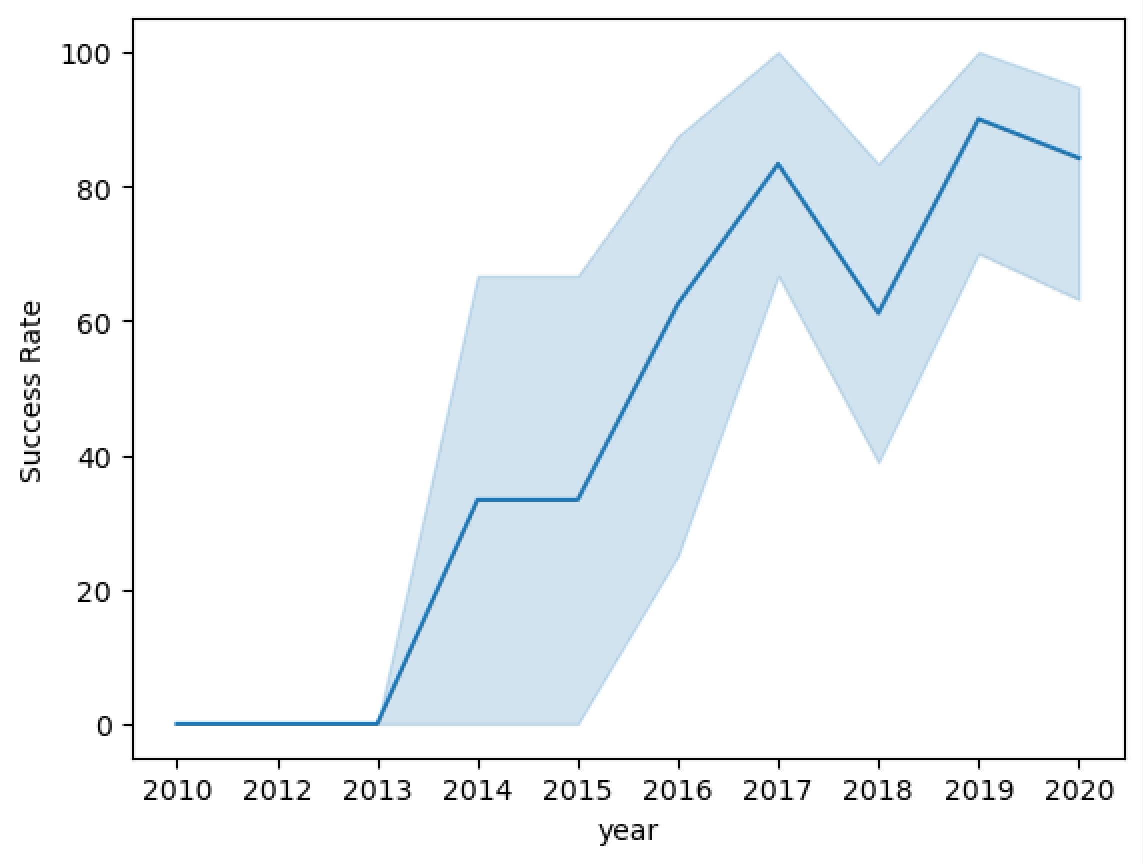


- Orange means successful launch
- Blue means unsuccessful launch
- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Launch Success Yearly Trend

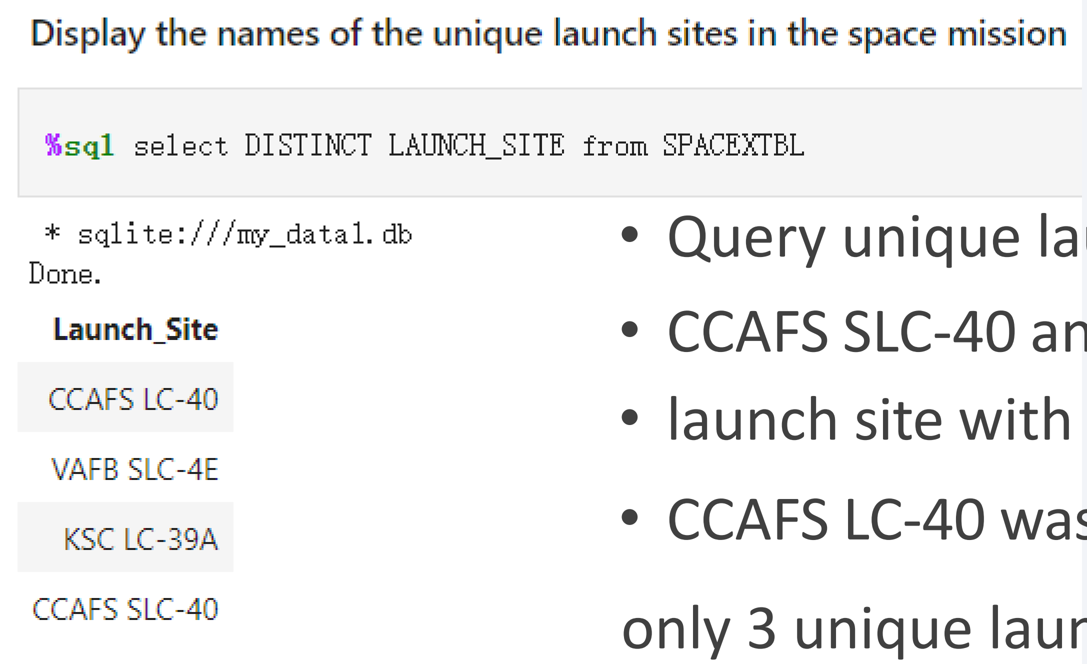
- Success generally increases over time from 2013 with a slight drop in 2018

- Success in recent years at around 80%



- 95% confidence interval

All Launch Site Names



Display the names of the unique launch sites in the space mission

```
%sql select DISTINCT LAUNCH_SITE from SPACEXTBL
```

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

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

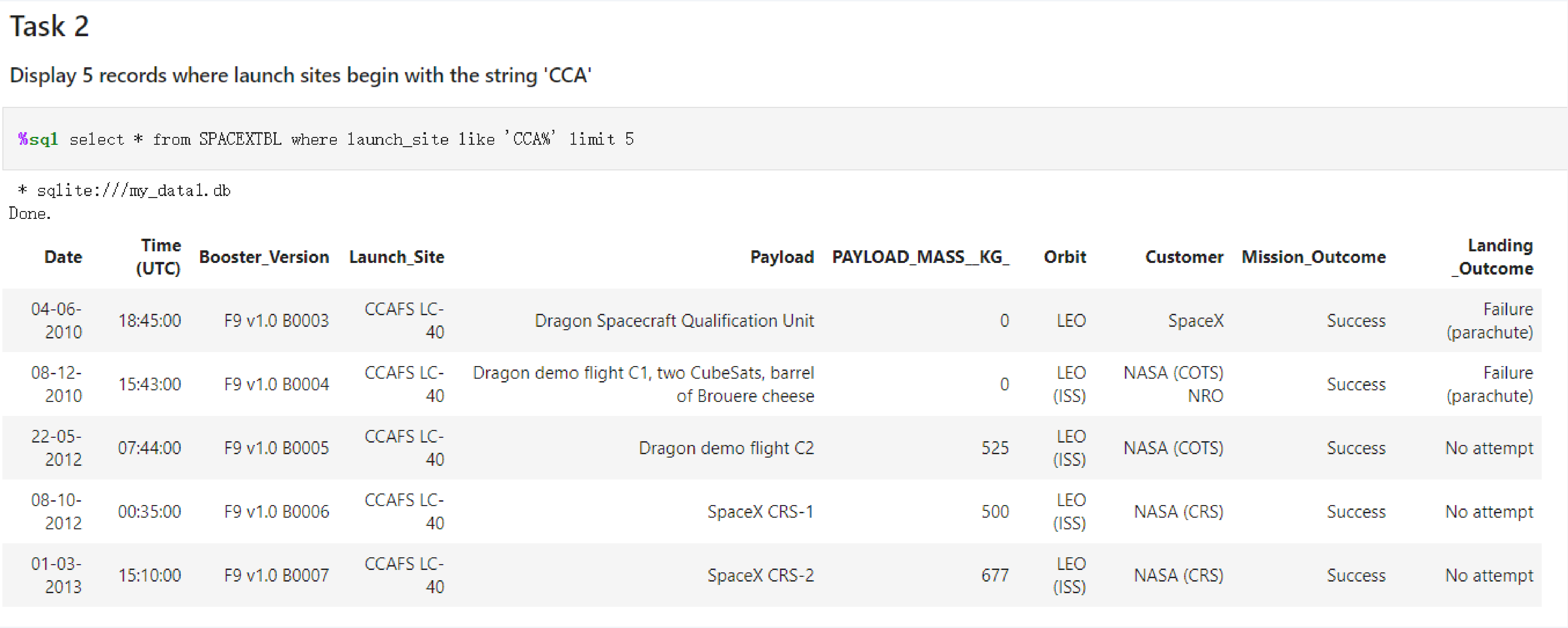
- Query unique launch site names from database.
- CCAFS SLC-40 and CCAFSSLC-40 likely all represent the same
- launch site with data entry errors.
- CCAFS LC-40 was the previous name. Likely

only 3 unique launch_site values: CCAFS SLC-40,

KSC LC-39A, VAFB SLC-4E

Launch Site Names Begin with 'CCA'

- First five entries in database with Launch Site name beginning with CCA.



Task 2

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

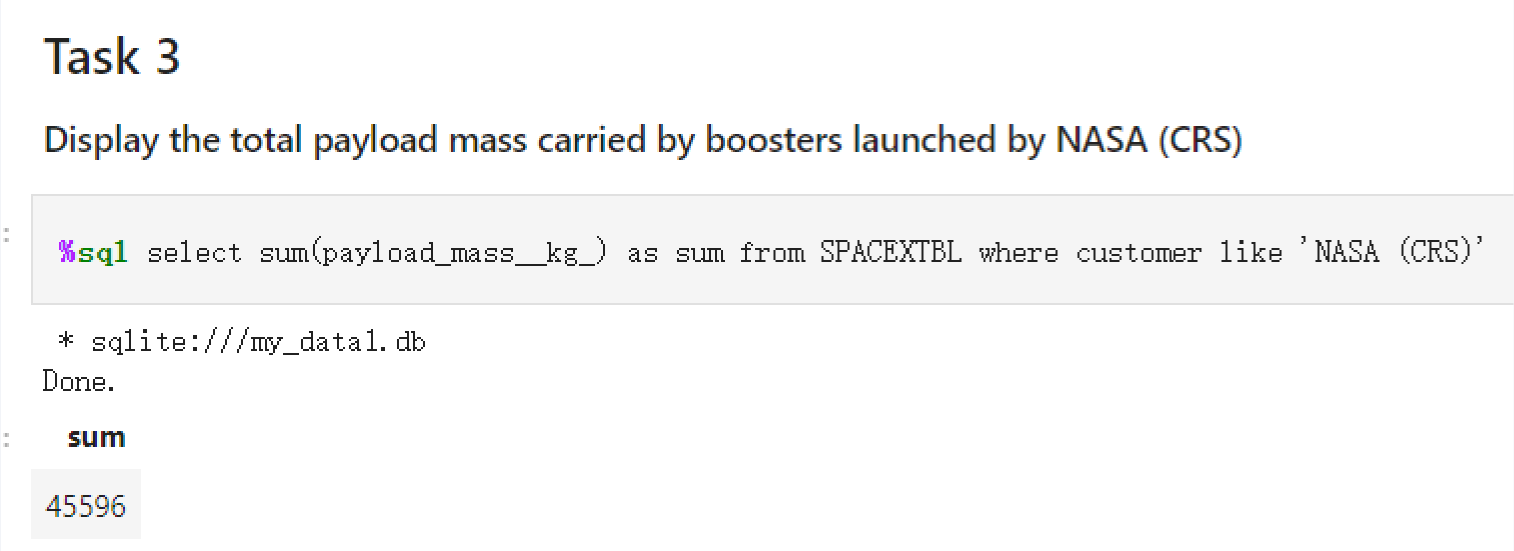
```
%sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5
```

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

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

Total Payload Mass

- This query sums the total payload mass in kg where NASA was the customer.
- CRS stands for Commercial Resupply Services which indicates that these payloads were sent to the International Space Station (ISS).



Task 3

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

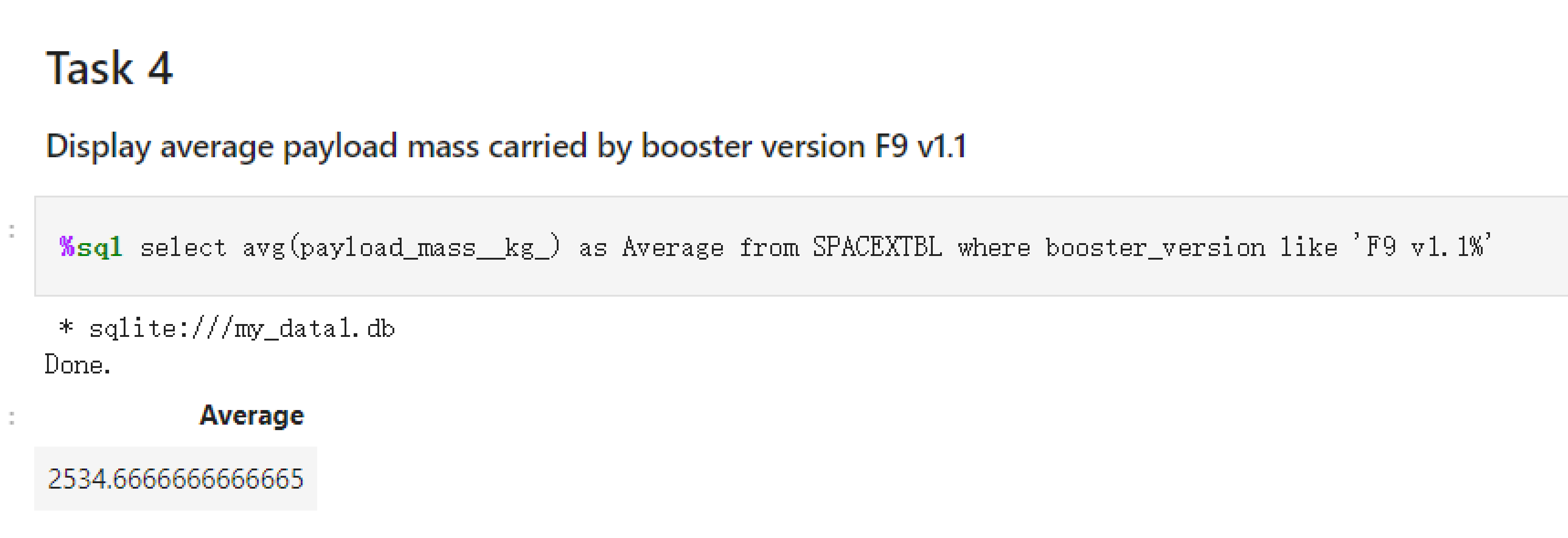
```
%sql select sum(payload_mass__kg_) as sum from SPACEXTBL where customer like 'NASA (CRS)'
```

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

sum
45596

Average Payload Mass by F9 v1.1

- This query calculates the average payload mass or launches which used booster version F9 v1.1
- Average payload mass of F9 1.1 is on the low end of our payload mass range



Task 4

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

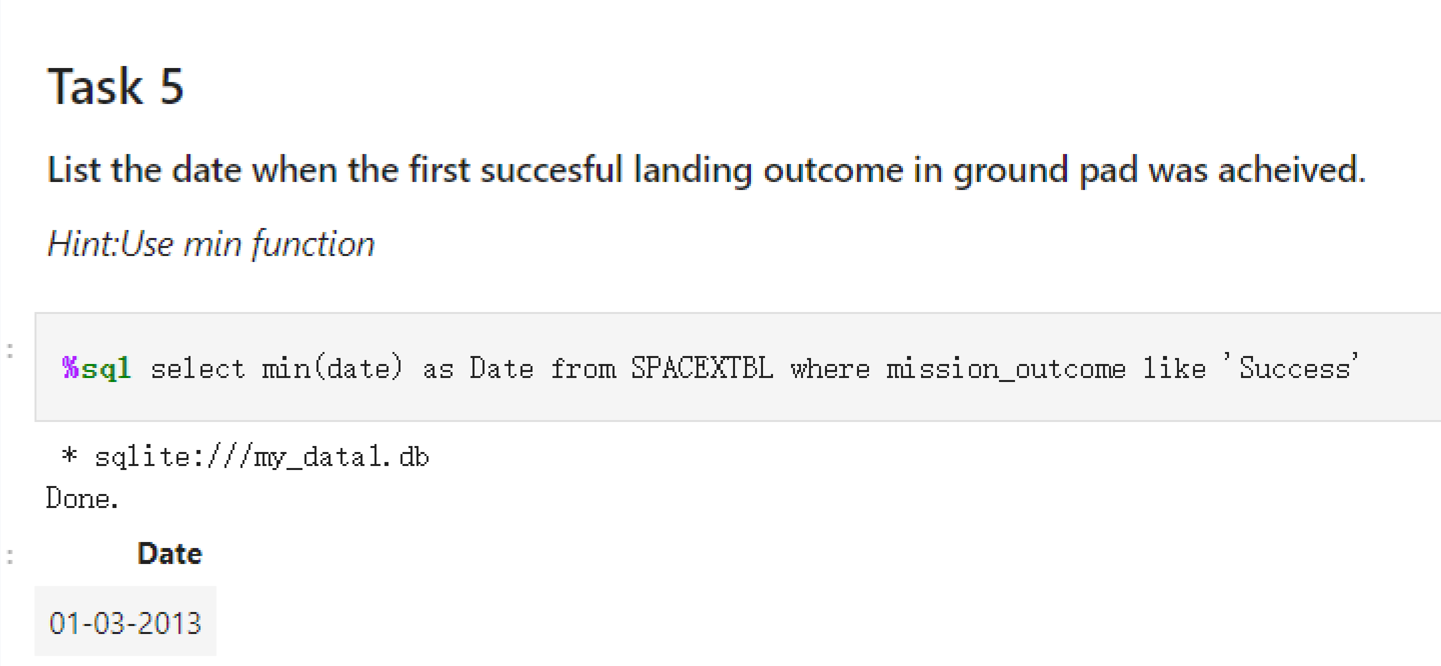
```
%sql select avg(payload_mass__kg_) as Average from SPACEXTBL where booster_version like 'F9 v1.1%'
```

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

Average
2534.6666666666665

First Successful Ground Landing Date

- This query returns the first successful ground pad landing date.



Task 5

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

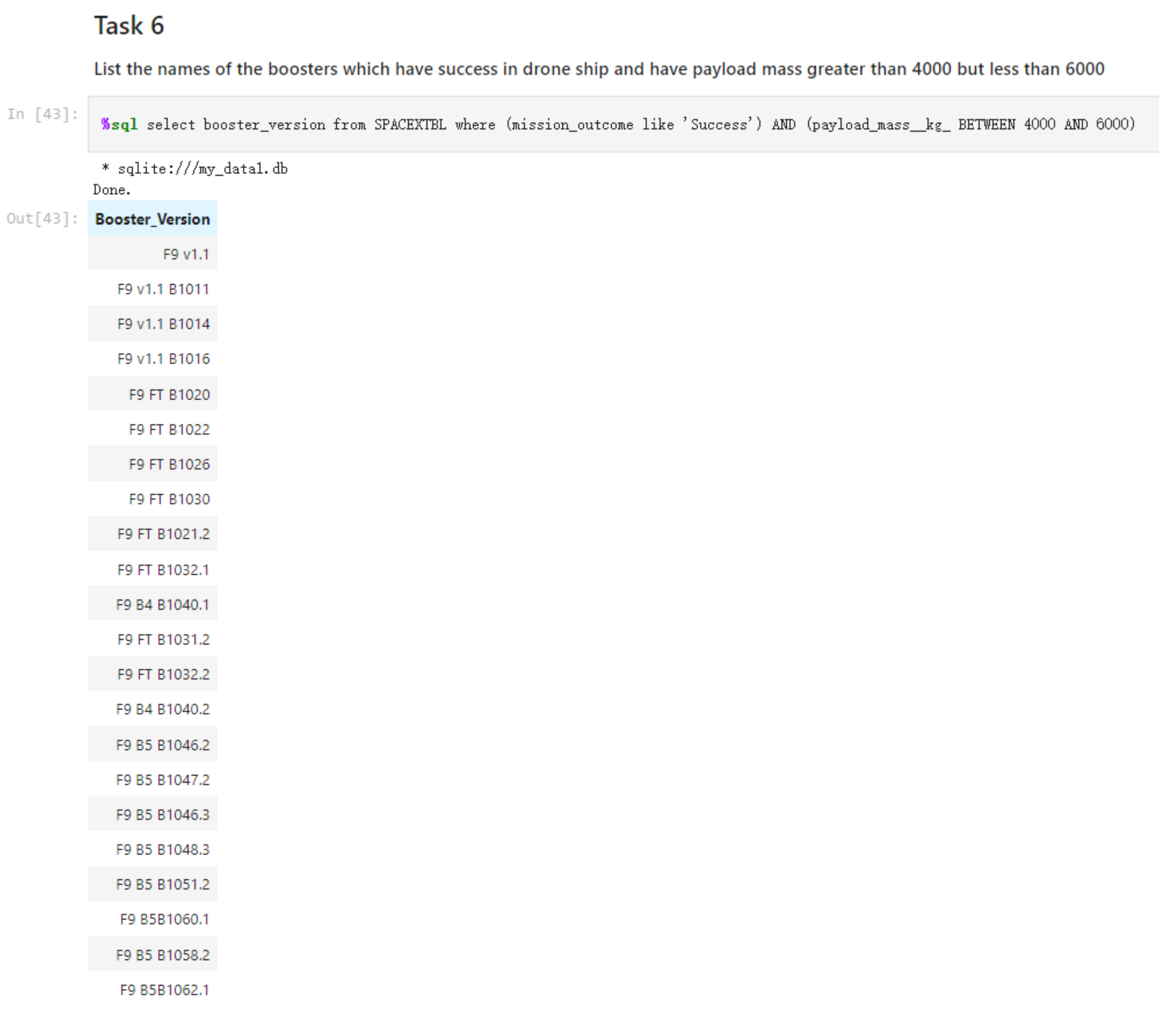
Hint:Use min function

```
%sql select min(date) as Date from SPACEXTBL where mission_outcome like 'Success'
```

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

Date
01-03-2013

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

```
In [43]: %sql select booster_version from SPACEXTBL where (mission_outcome like 'Success') AND (payload_mass__kg_ BETWEEN 4000 AND 6000)
```

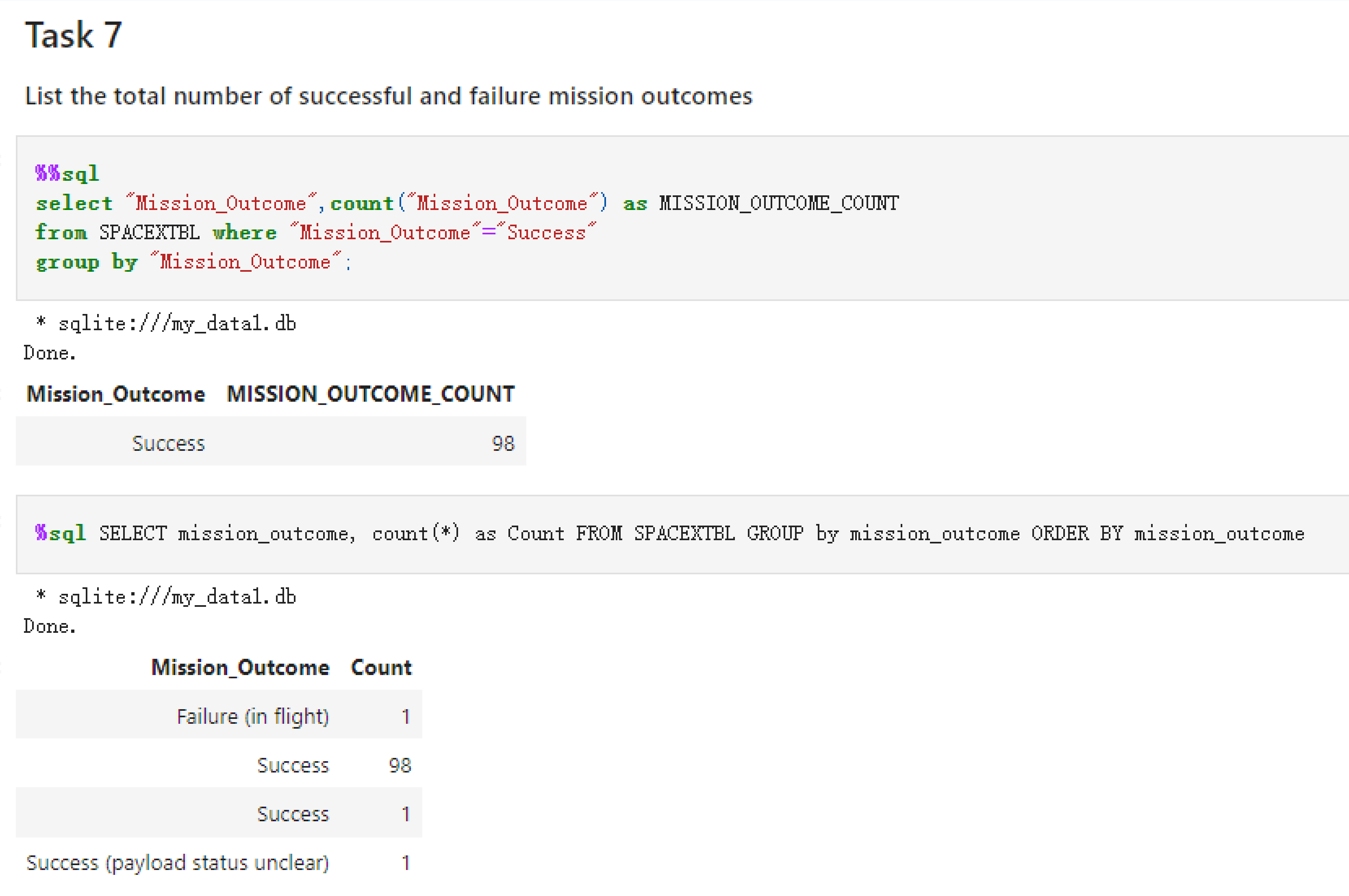
```
 * sqlite:///my_data1.db
Done.
```

Out[43]:

Booster_Version
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 FT B1032.2
F9 B4 B1040.2
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1046.3
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

- This query returns the all booster versions that had successful drone ship landings and a payload mass between 4000 and 6000 noninclusively.

Total Number of Successful and Failure Mission Outcomes



Task 7

List the total number of successful and failure mission outcomes

```
%%sql
select "Mission_Outcome",count("Mission_Outcome") as MISSION_OUTCOME_COUNT
from SPACEXTBL where "Mission_Outcome"="Success"
group by "Mission_Outcome";
```

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

Mission_Outcome	MISSION_OUTCOME_COUNT
Success	98

```
%sql SELECT mission_outcome, count(*) as Count FROM SPACEXTBL GROUP by mission_outcome ORDER BY mission_outcome
```

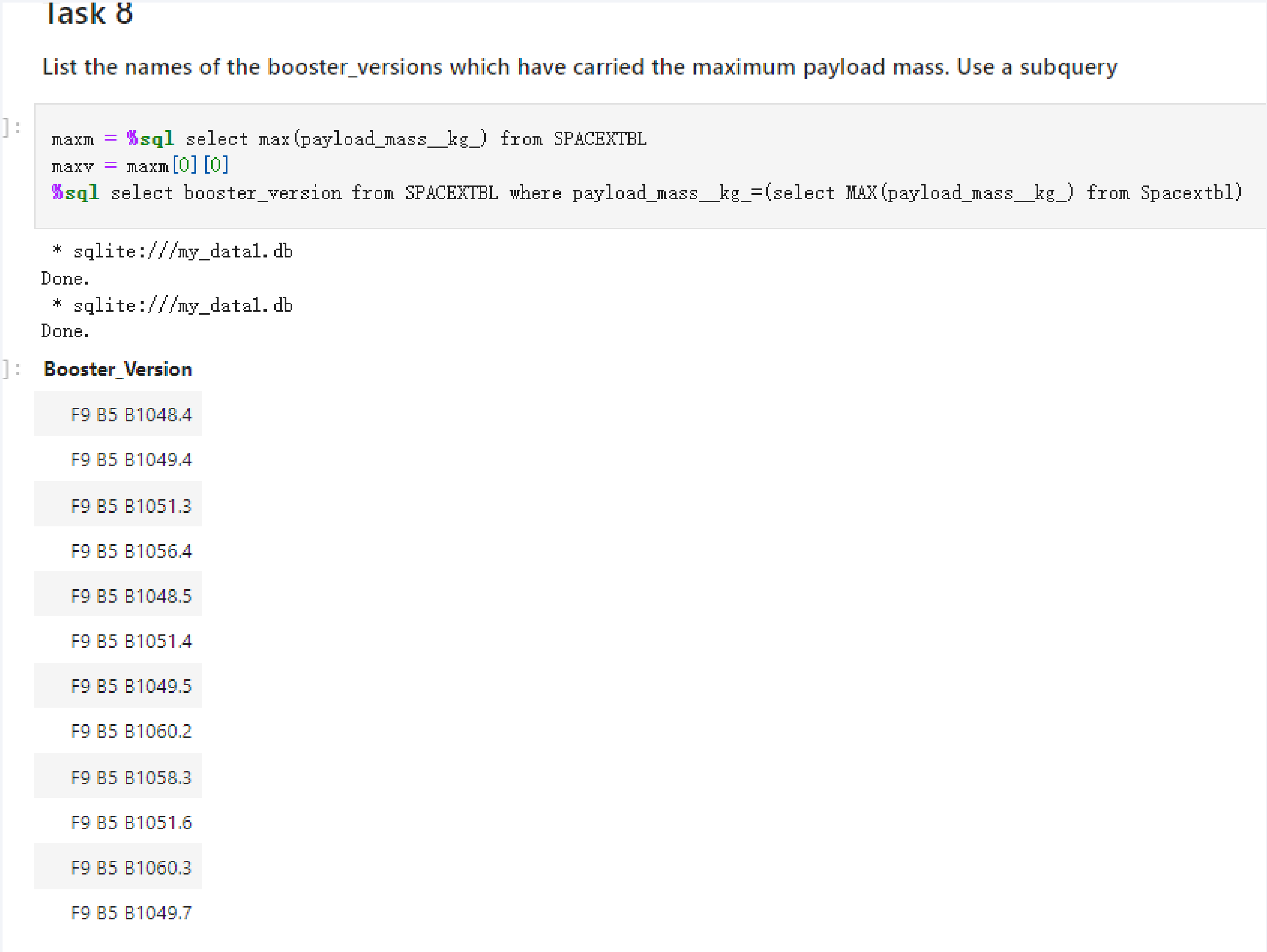
```
 * sqlite:///my_data1.db
Done.
```

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

- This query returns a count of each mission outcome.

- The mission is successful for 99%.

Boosters Carried Maximum Payload



Task 8

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

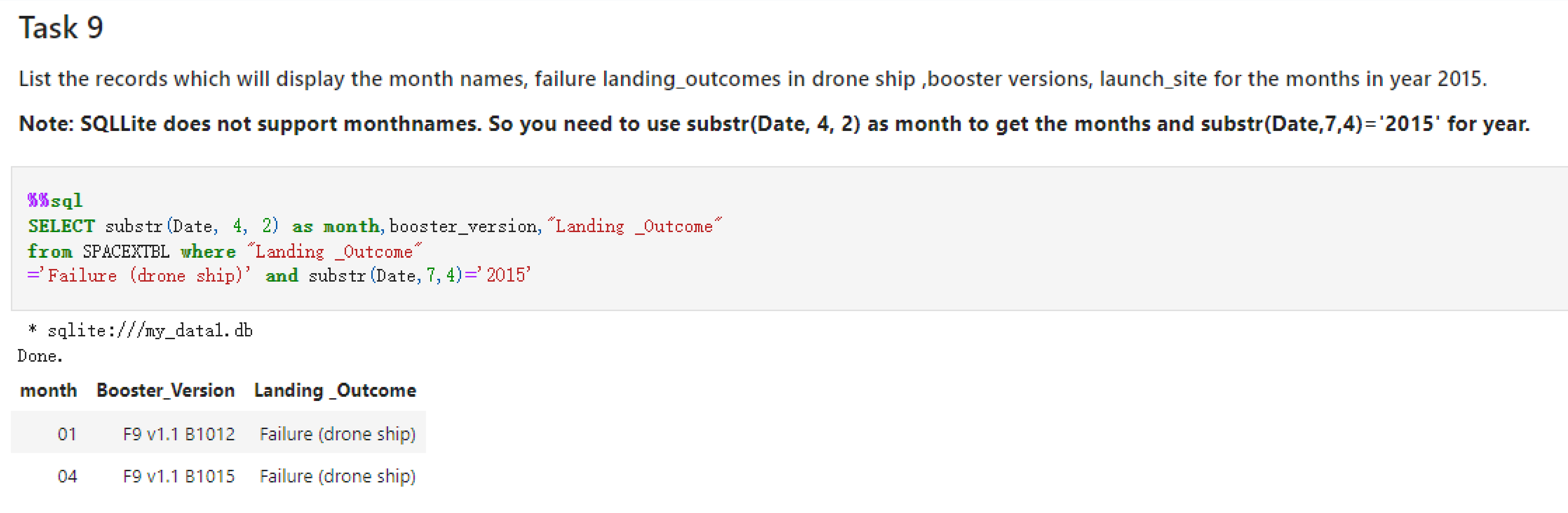
```
maxm = %sql select max(payload_mass__kg_) from SPACEXTBL
maxv = maxm[0][0]
%sql select booster_version from SPACEXTBL where payload_mass__kg_=(select MAX(payload_mass__kg_) from Spacextbl)
```

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

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

- This query returns the booster versions that carried the highest payload mass of 15600 kg.
- These booster versions are very similar and all are of the F9 B5 B10xx.x variety.
- This likely indicates payload mass correlates with the booster version that is used.

2015 Launch Records



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.

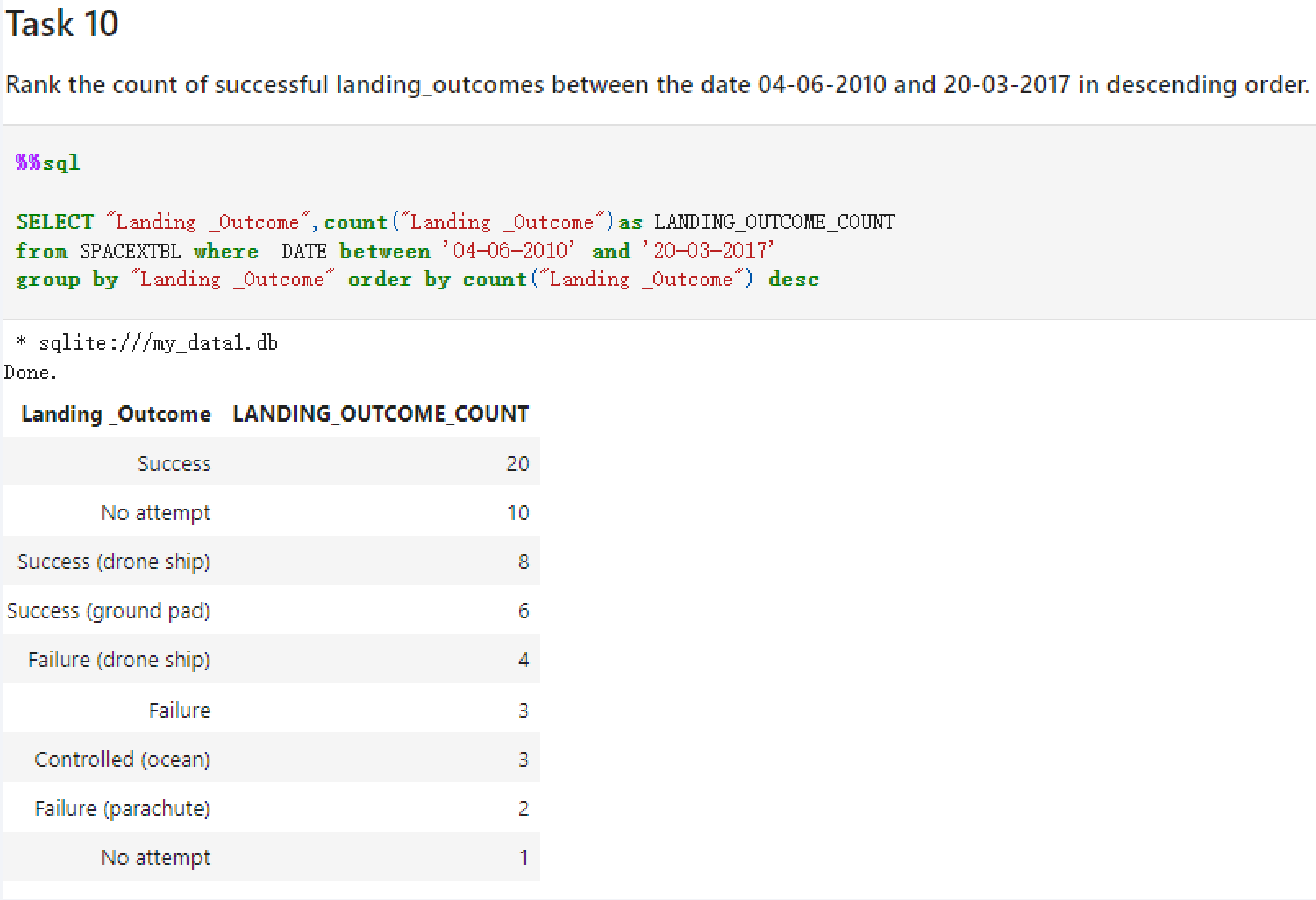
```
%%sql
SELECT substr(Date, 4, 2) as month,booster_version,″Landing _Outcome″
from SPACEXTBL where ″Landing _Outcome″
=’Failure (drone ship)’ and substr(Date,7,4)=’2015’
```

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

month	Booster_Version	Landing _Outcome
01	F9 v1.1 B1012	Failure (drone ship)
04	F9 v1.1 B1015	Failure (drone ship)

- This query returns the Month, Landing Outcome, Booster Version, Payload Mass (kg), and Launch site of 2015 launches where stage 1 failed to land on a drone ship.
- There were two such occurrences.

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



Task 10

Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

```
%%sql

SELECT "Landing _Outcome",count("Landing _Outcome")as LANDING_OUTCOME_COUNT
from SPACEXTBL where  DATE between '04-06-2010' and '20-03-2017'
group by "Landing _Outcome" order by count("Landing _Outcome") desc
```

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

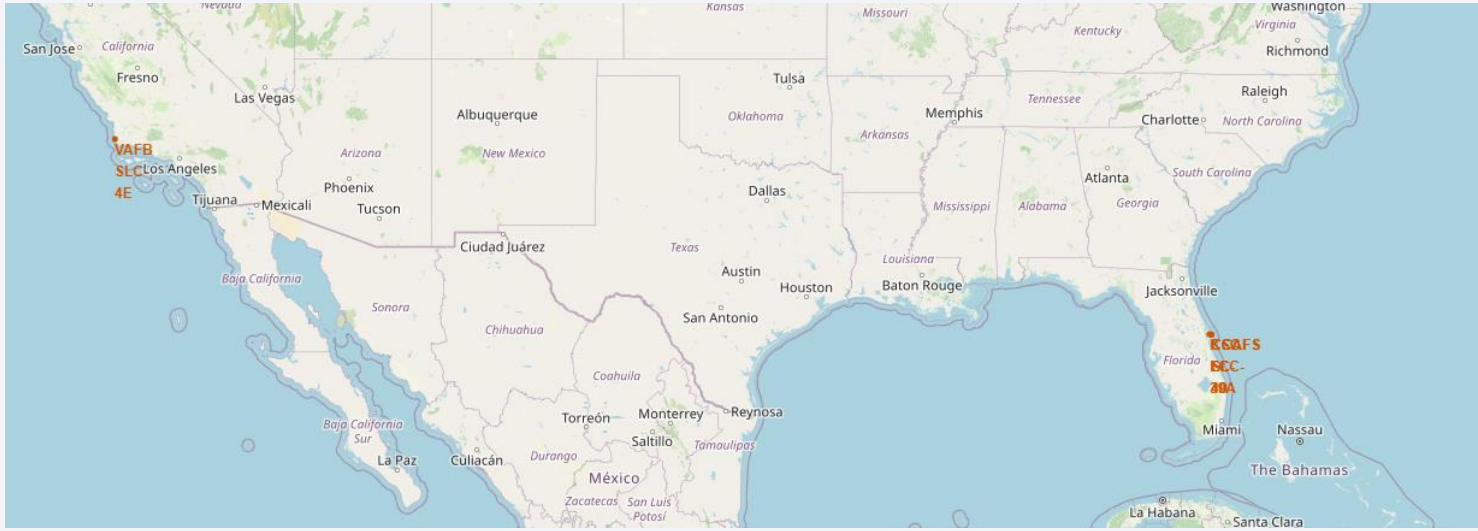
Landing _Outcome	LANDING_OUTCOME_COUNT
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

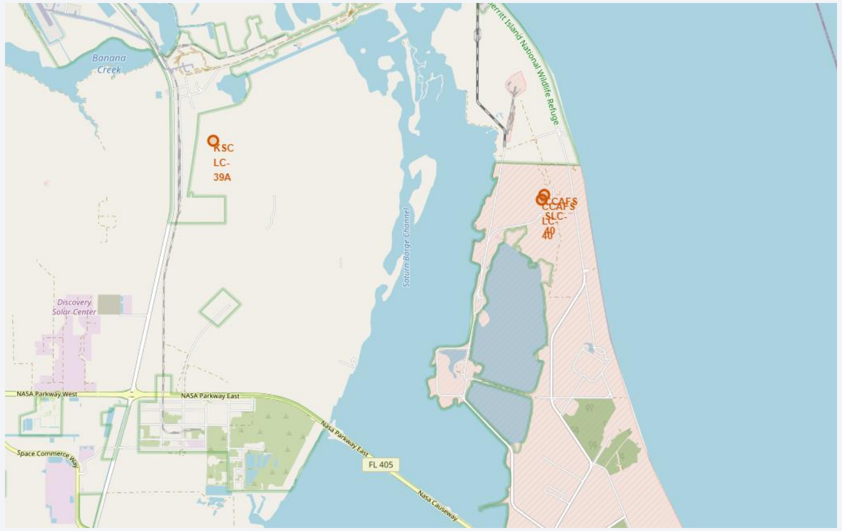
- This query returns a list of successful landings and between 2010-06-04 and 2017-03-20 inclusively.
- There are two types of successful landing outcomes: drone ship and ground pad landings.
- There were 8 successful landings in total during this time period

Section 3

Launch Sites Proximities Analysis

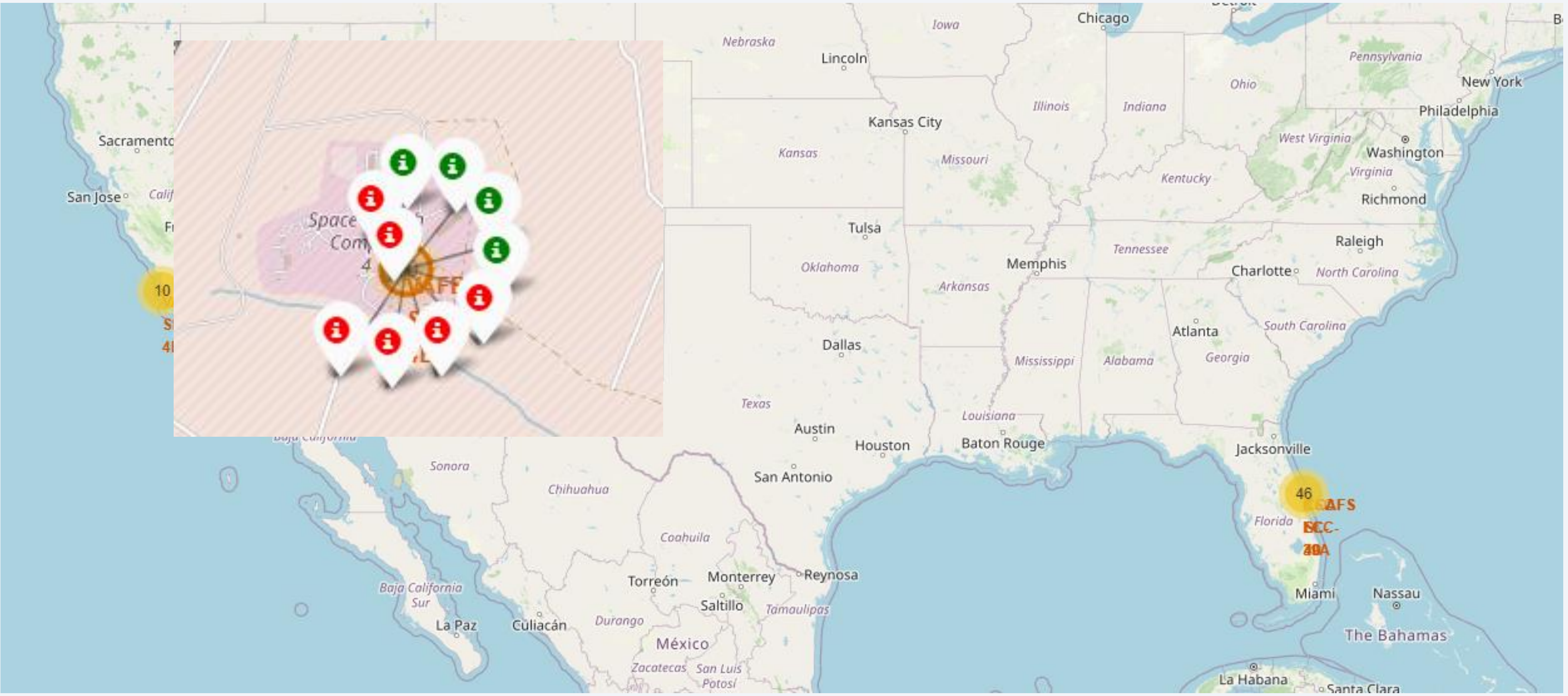
Launch Site Locations





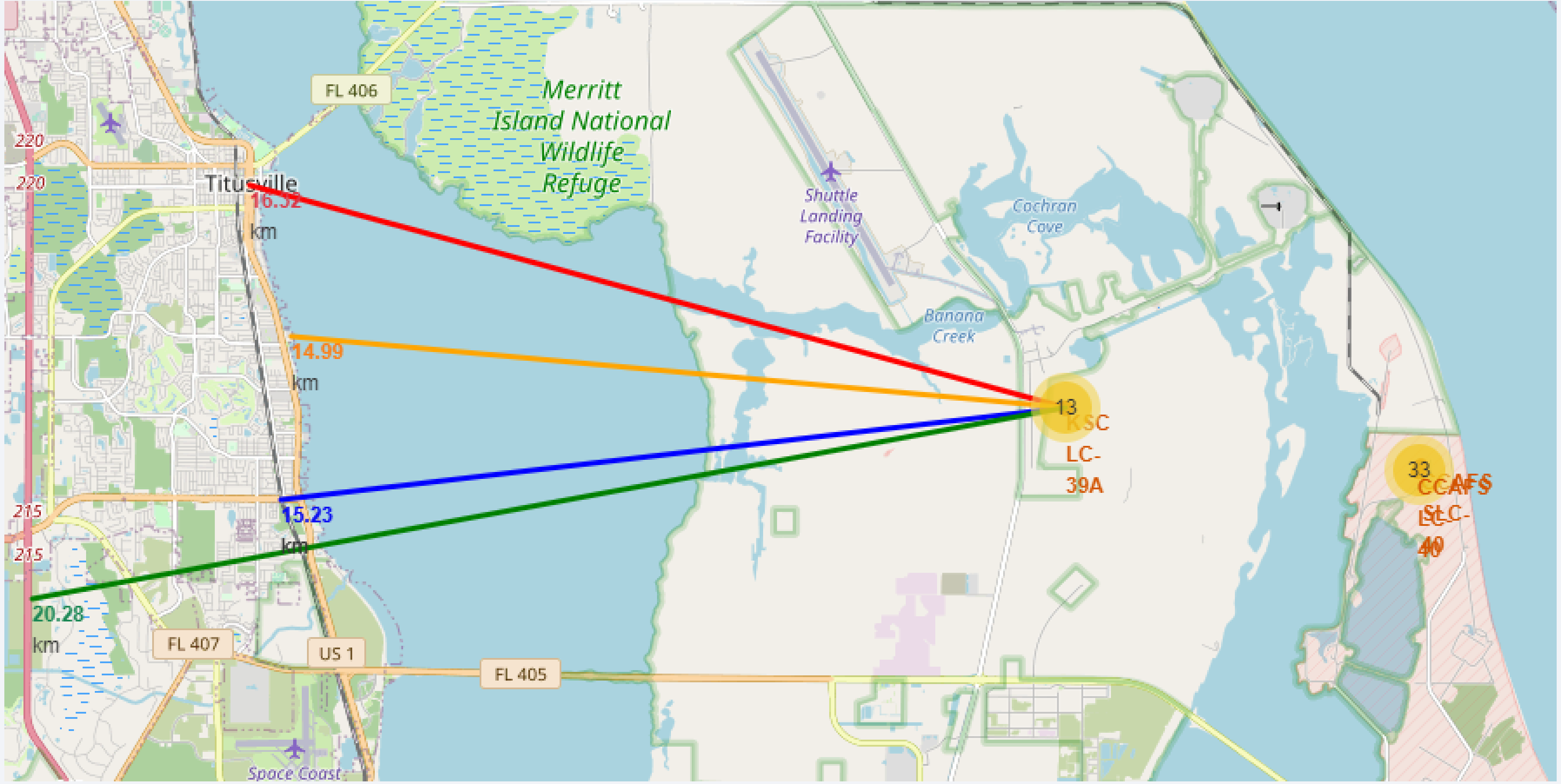
- The left map shows all launch sites relative US map.
- The right map shows the two Florida launch sites since they are very close to each other.
- All launch sites are near the ocean.

Color-Coded Launch Markers



- Clusters on Folium map can be clicked on to display each successful landing (green icon) and failed
- landing (red icon). In this example VAFB SLC-4E shows 4 successful landings and 6 failed landings.

Key Location Proximities



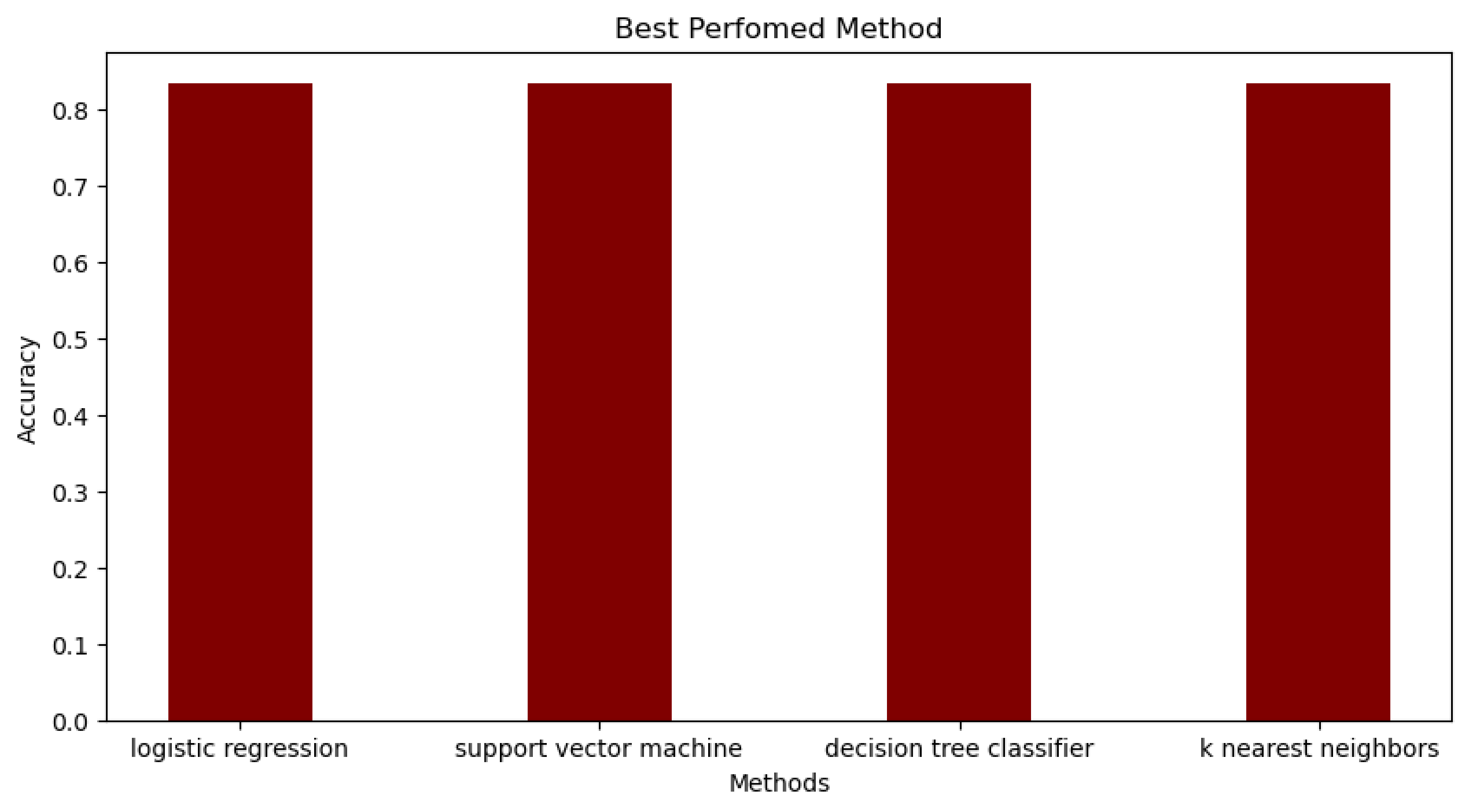
- The lines show the different distances between KSC LC-39A and other places

Section 5

Predictive Analysis (Classification)

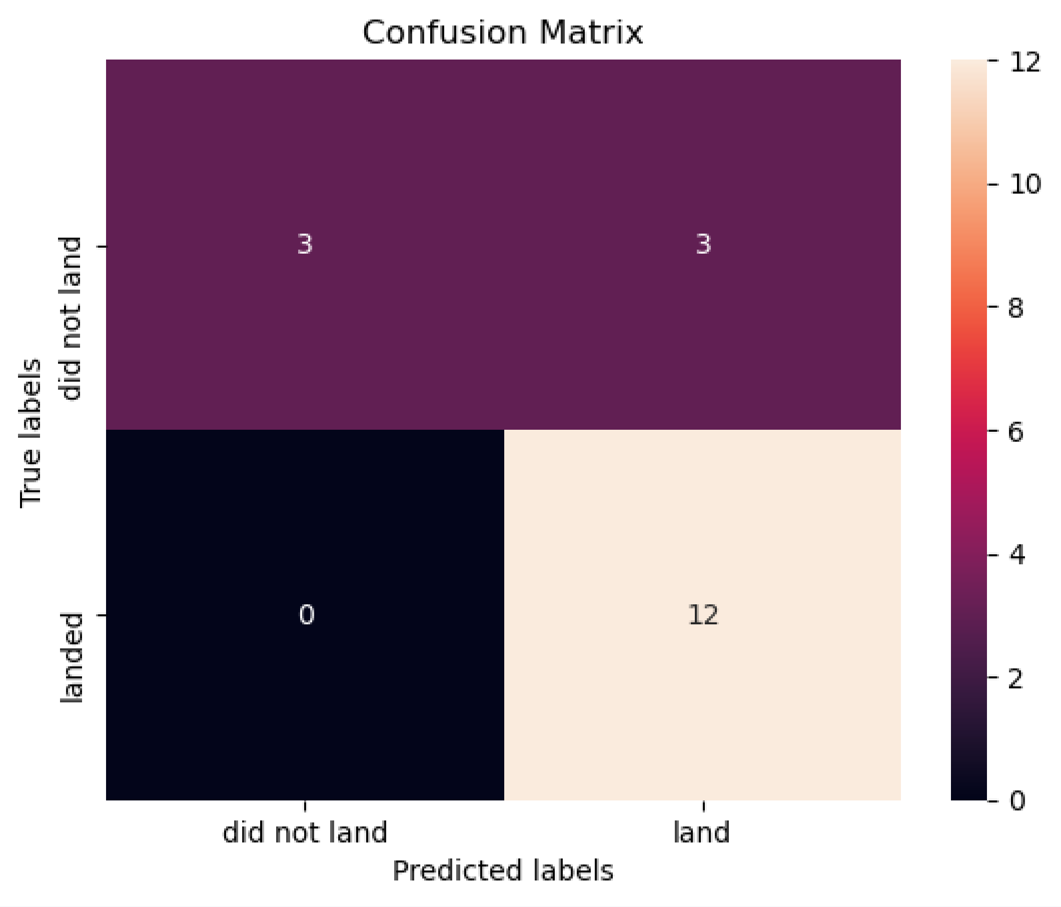
Classification Accuracy

- All models had virtually the same accuracy on the test set at 83.33% accuracy.
- It should be noted that test size is small at only sample size of 18.



Confusion Matrix

- Since all models performed the same for the test set, the confusion matrix is the same across all models.
- The models predicted 12 successful landings when the true label was successful landing.
- The models predicted 3 unsuccessful landings when the true label was unsuccessful landing.
- The models predicted 3 successful landings when the true label was unsuccessful landings (false positives). Our models over predict successful landings.



Conclusions

- Task: to develop a machine learning model for Space Y bidding against SpaceX
- The goal of model is to predict when Stage 1 will successfully land to save ~$100 million USD
- Used data from a public SpaceX API and web scraping SpaceX Wikipedia page
- Created data labels and stored data into a DB2 SQL database
- Created a dashboard for visualization
- We created a machine learning model with an accuracy of 83%
- Allon Mask of SpaceY can use this model to predict with relatively high accuracy whether a launch will have a successful Stage 1 landing before launch to determine whether the launch should be made or not
- If possible more data should be collected to better determine the best machine learning model and improve accuracy

Thank you!