EVERYDAY ASTRONAUT

APPLIED DATA SCIENCE CAPSTONE

An Analyst on Launch Data of SpaceX

By Mehmet Cem Öztürk

OUTLINE

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

EXECUTIVE SUMMARY

Data acquisition was done through web scraping and SpaceX API. Exploratory data analysis (EDA), including data processing, data visualization and interactive visual analysis, machine learning prediction. Summary of all results. I collected valuable data from public sources. EDA allowed to identify the best features for predicting the success of launches. Prediction by machine learning showed the best model for predicting which features are important to make the most of the opportunity, using all the data collected.

INTRODUCTION

The task is to produce an analysis, based on data, that best estimates the total cost of launches, by comparing the successful landings of thefirst stage of rockets. To represent these with graphs and to be able to find out a possible optimal rocket model, landing site.

METHODOLOGY

- Space X data were obtained from sources:Space X API (https://api.spacexdata.com/v4/rockets/)
- WebScraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)
- The collected data was enriched by creating a landing result label based on the result data after summarizing and analyzing features.
- Performing exploratory data analysis (EDA) using visualization and SQL.
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

DATA COLLECTION – SPACEX API

1. Request API and parse the SpaceX launch data
2. Filter data to only include Falcon 9 launches
3. Deal with Missing Values
4. Replace missing PayloadMass values with mean

DATA COLLECTION WEB-SCRAPING

1. Request the Falcon9 Launch on Wikipedia
2. Extract all column/variable names from the HTML table header
3. Create a data frame by parsing the launch HTML tables
4. Iterate through table cells to extract data to dictionary
5. Create dictionary

DATA WRANGLING

First, an exploratory data analysis (EDA) of the dataset was performed, then the launch summaries per site, the frequency of each orbit, and the frequency of mission results per orbit type were calculated. Lastly, the landing result designation was created from the "Result" column.

EDA WITH DATA VISUALIZATION

To explore data, scatterplots and barplots were used to visualize the relationship between pair of features:Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass, Orbit and Flight Number, Payload and Orbit

EDA WITH SQL

1. Names of all launch sites in the space mission SQL queries were performed: Top 5 launch sites whose name begins with the term "CCA". Total Payload Mass of NASA Launched Boosters (CRS).
2. The average payload mass of the booster version F9 v1.1.
3. Date of first successful ground landing. Names of boosters that successfully landed in the drone ship and have a payload mass between 4000 and 6000 kg. Number of successful and failed mission outcomes.
4. Names of booster versions that carried the maximum payload mass;Failed landings in drone ships, their booster versions and the names of the launch sites for 2015, also a ranking of the number of landing results between 04/06/2010 and 20/03/2017.

BUILD AN INTERACTIVE MAP WITH FOLIUM

Markers indicate points such as launch sites. Circles indicate highlighted areas around specific coordinates, such as NASA Johnson Space Center Marker clusters indicate groups of events at each coordinate, such as launches at a launch site; and Lines are used to indicate distances between two coordinates.

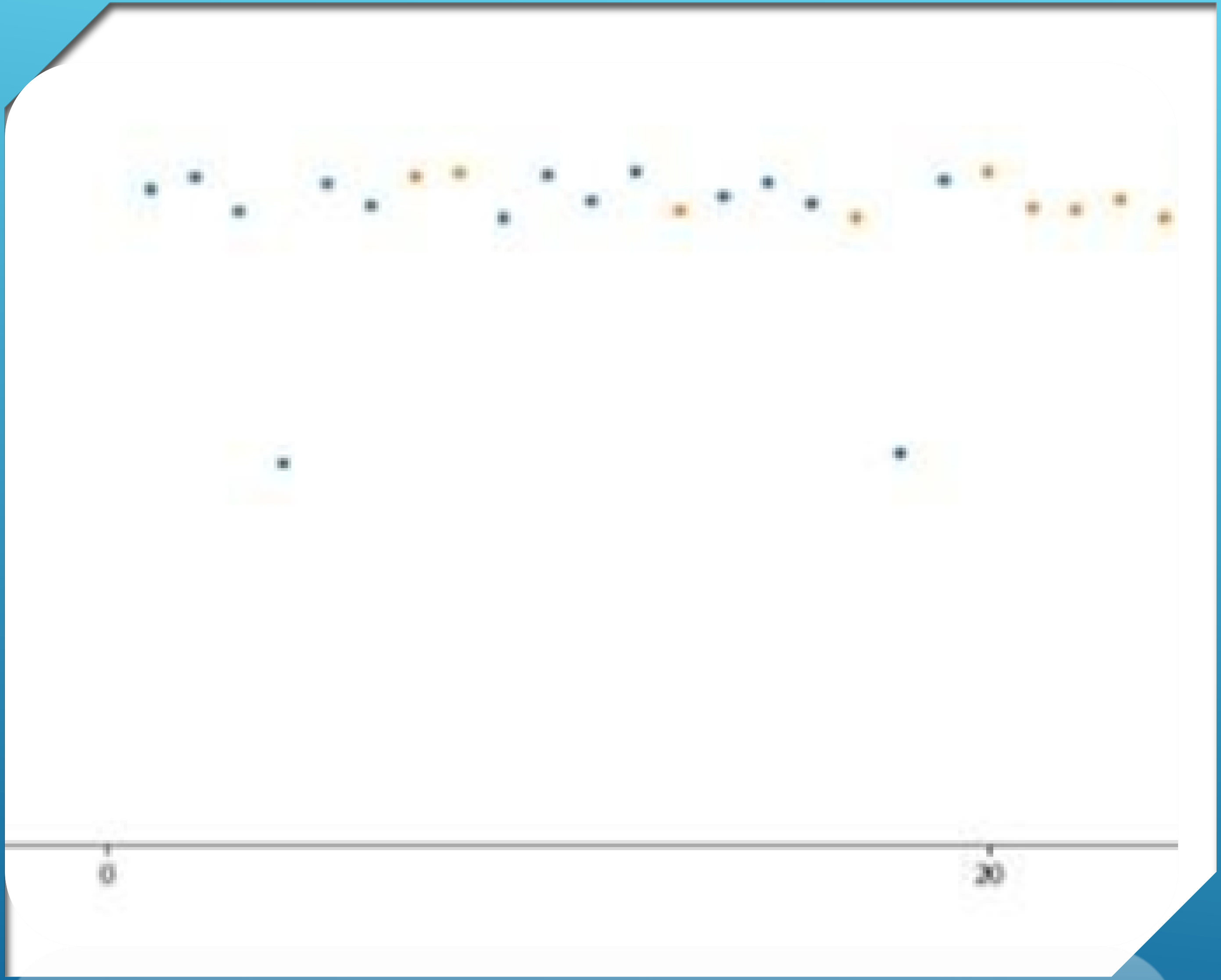
PREDICTIVE ANALYSIS

1. Data preparation and standardization
2. Test of each model with combinations of hyperparameters
3. Comparison of results

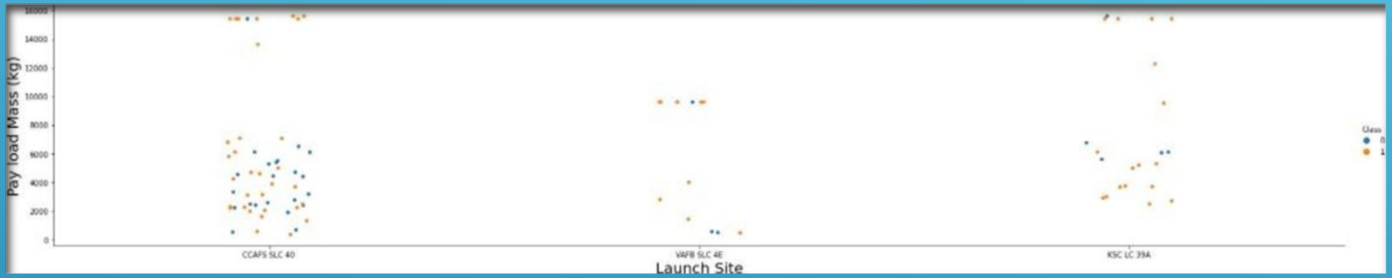
❖ RESULTS:

E.D.A WITH VISUALATION
E.D.A WITH SQL
INTERACTIVE FORUM MAPS
BUILD A DASHBOARD WITH PLOTLY DASH
PREDICTIVE ANALYSES

FLIGHT NUMBER VS. LAUNCHSITE

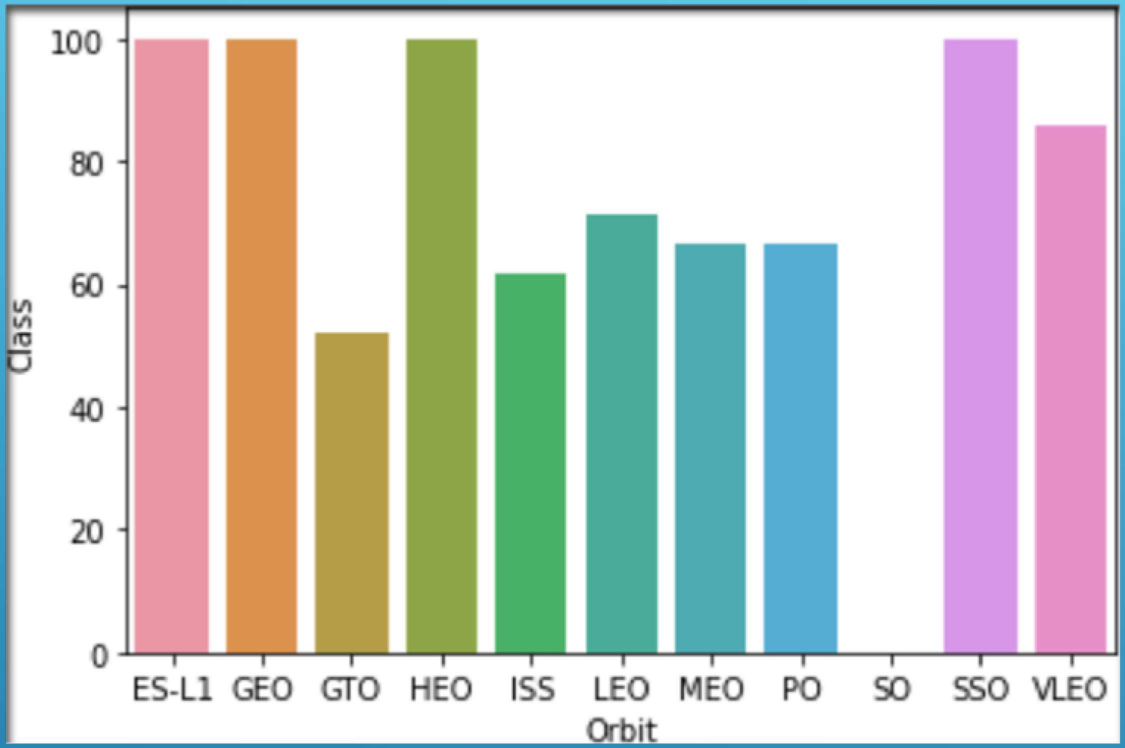


- Graphic suggests an increase in success rate over time (indicated in Flight Number). Likely a big breakthrough around flight 20 which significantly increased success rate. CCAFS appears to be the main launch site as it has the most volume.



- Payloadmass appears to fall mostly between 0-7000 kg. Different launch sites also seem to use different payload mass.

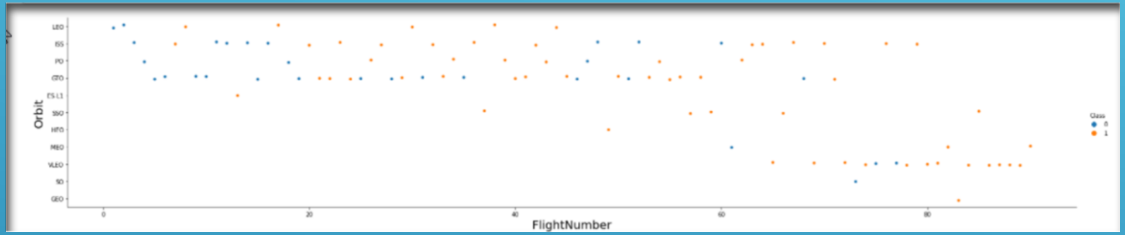
PAYLOAD VS. LAUNCHSITE



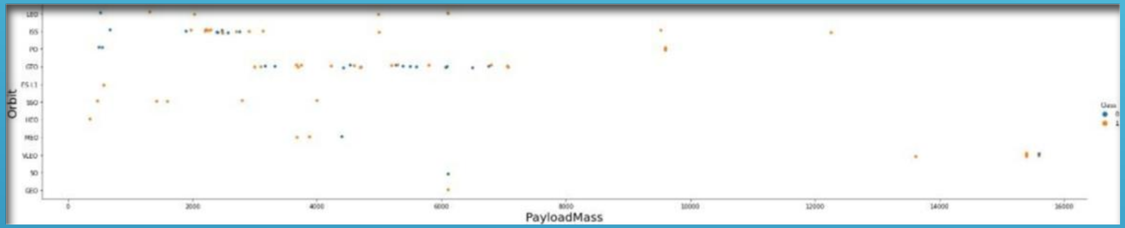
- ES-L1 (1), GEO (1), HEO (1) have 100% success rate (sample sizes in parenthesis) SSO (5) has 100% success rate
- VLEO (14) has decent success rate and attempts SO (1) has 0% success rate
- GTO (27) has the around 50% success rate but largest sample

SUCCESSRATE VS. ORBITTYPE

FLIGHT NUMBER VS. ORBITTYPE

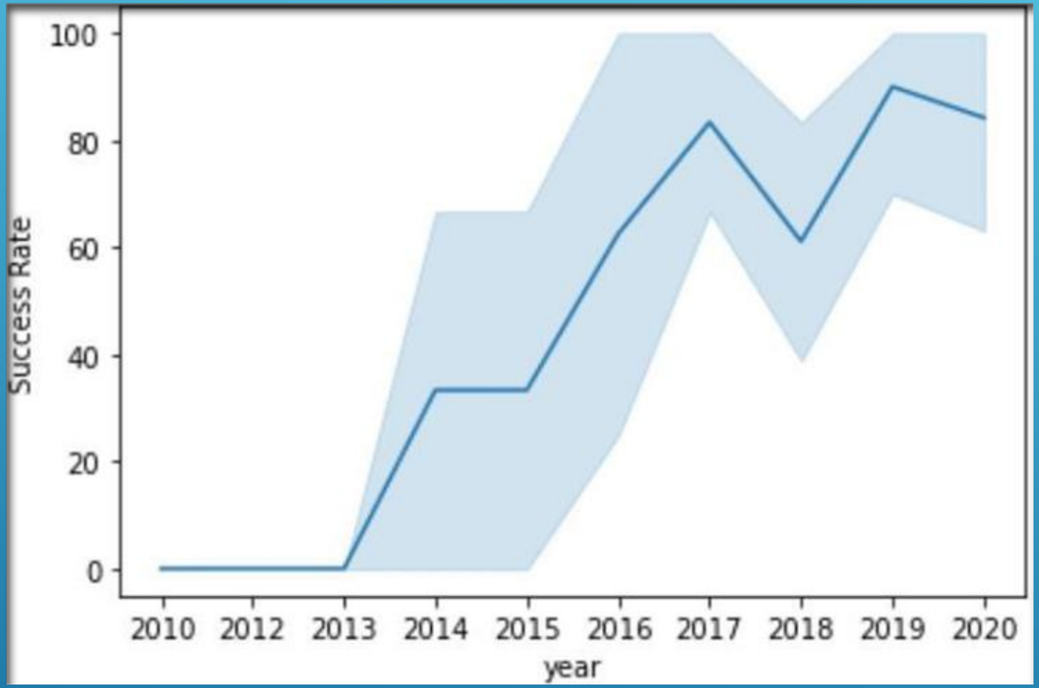


- Launch Orbit preferences changed over Flight Number. Launch Outcome seems to correlatewith this preference.
- SpaceX started with LEO orbits which saw moderate success LEO and returned to VLEO in recent launches
- SpaceX appears to perform better in lower orbits or Sun-synchronous orbits



- Payload mass seems to correlatewith orbit
- LEO and SSO seem to have relatively low payload mass
- The other most successful orbit VLEO only has payload mass values in the higher end of the range

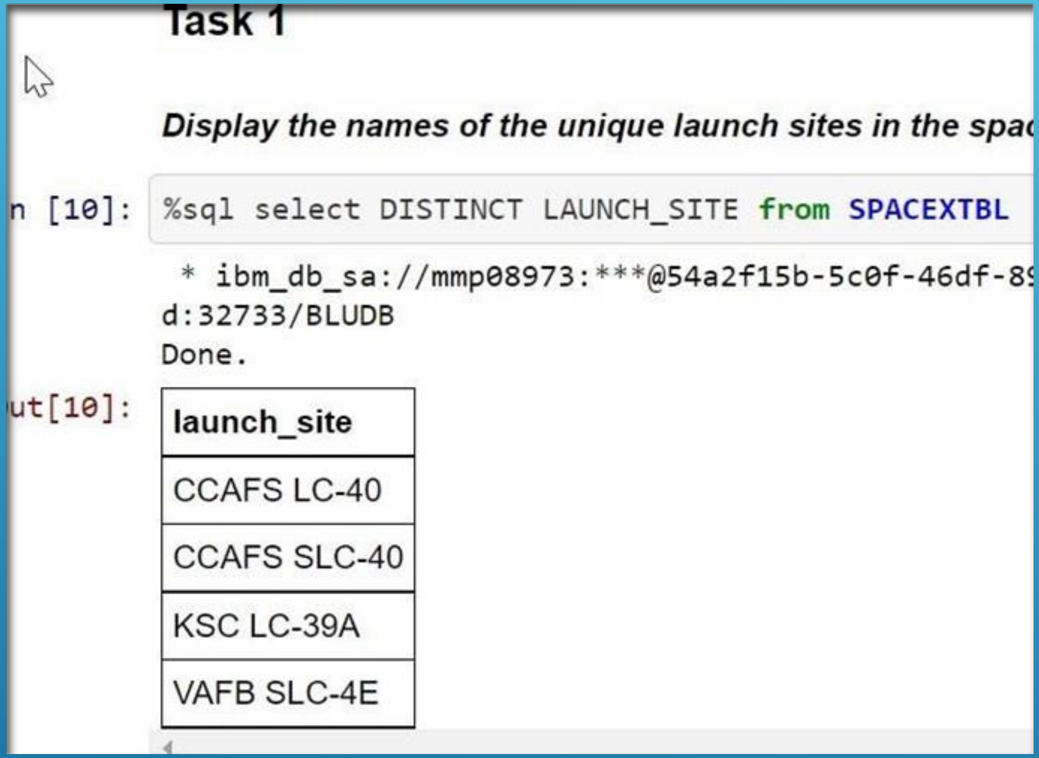
PAYLOAD VS. ORBIT TYPE



- Success generally increases over time since 2013 with a slight dip in 2018 Success in recent years at around 80%

LAUNCH SUCCESSYEARLYTREND

ALL LAUNCH SITE NAMES



Task 1

Display the names of the unique launch sites in the spac

```
n [10]: %sql select DISTINCT LAUNCH_SITE from SPACEXTBL
```

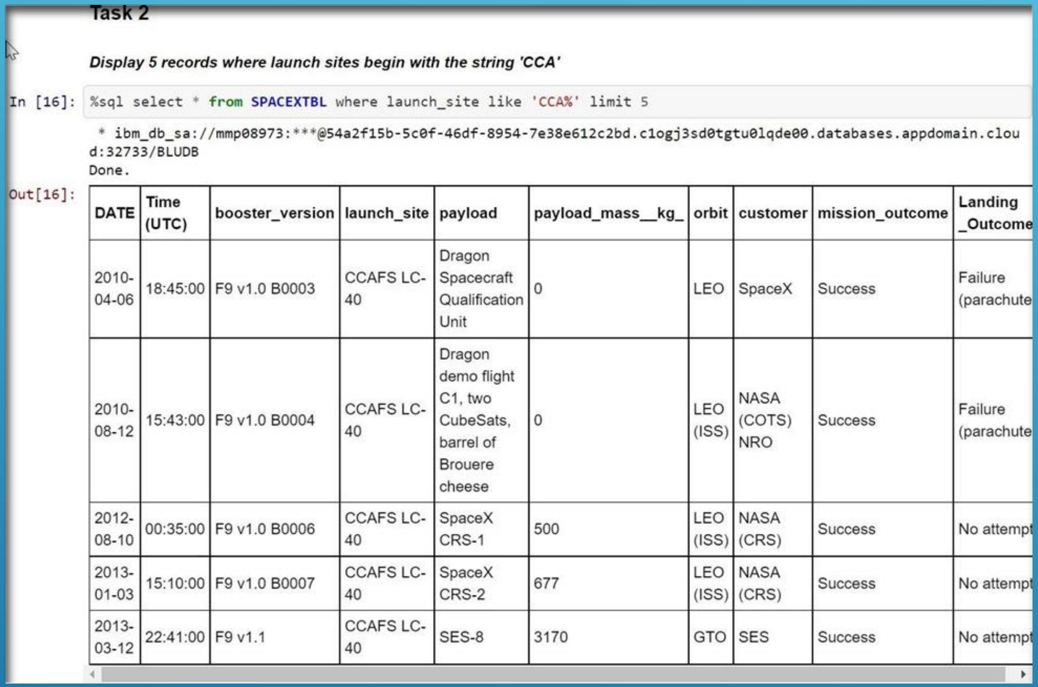
```
 * ibm_db_sa://mmp08973:***@54a2f15b-5c0f-46df-89
d:32733/BLUDB
Done.
```

ut[10]:

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

They are obtained by selecting unique occurrences of “launch_site” values

LAUNCH SITE NAMESBEGINNING WITH `CCA`



Task 2

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

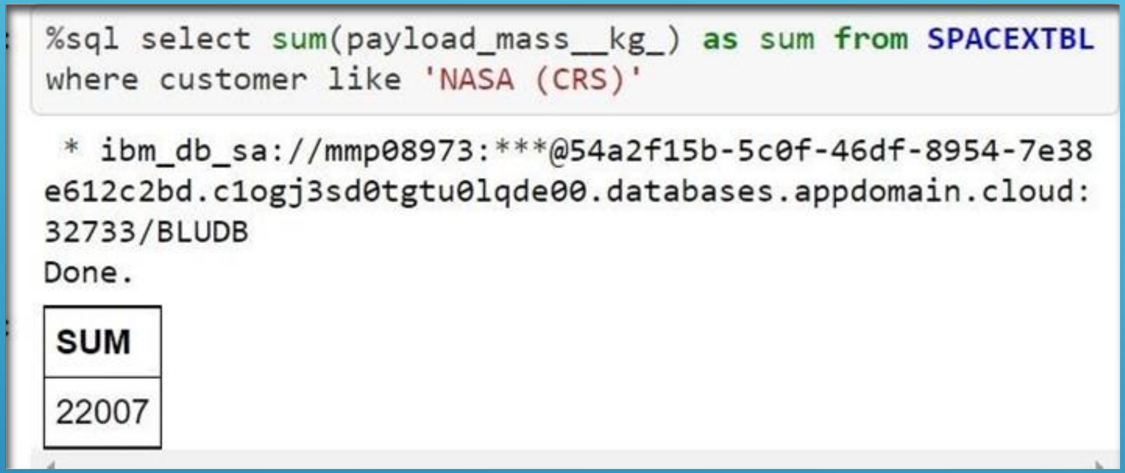
```
In [16]: %sql select * from SPACEXTBL where launch_site like 'CCA%' limit 5

 * ibm_db_sa://mmp08973:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.clou
d:32733/BLUDB
Done.
```

Out[16]:

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
2012-08-10	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
2013-03-12	22:41:00	F9 v1.1	CCAFS LC-40	SES-8	3170	GTO	SES	Success	No attempt

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

TOTAL PAYLOAD MASS FROM NASA

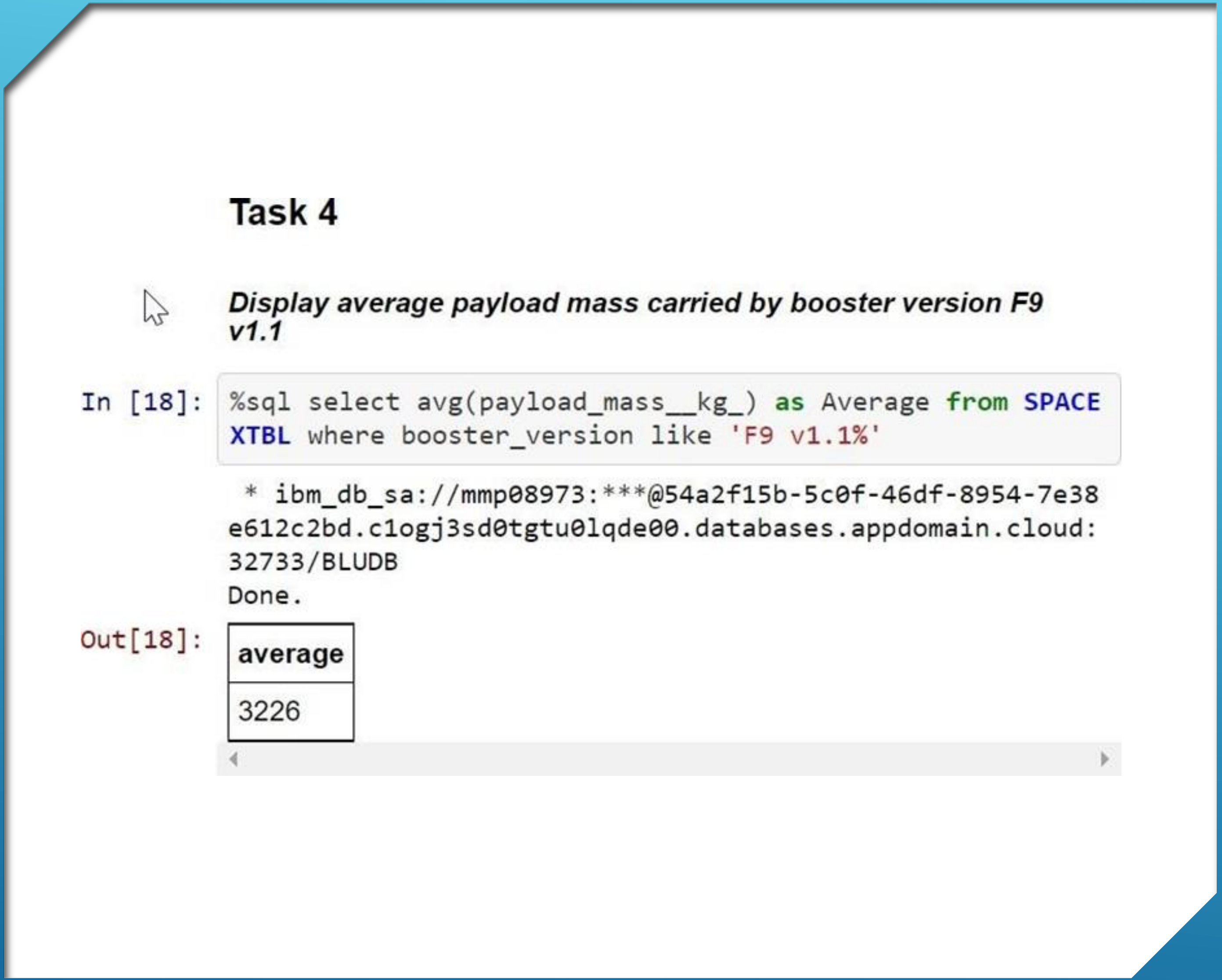


```
%sql select sum(payload_mass__kg_) as sum from SPACEXTBL
where customer like 'NASA (CRS)'
```

```
 * ibm_db_sa://mmp08973:***@54a2f15b-5c0f-46df-8954-7e38
e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:
32733/BLUDB
Done.
```

SUM
22007

- This query sums the totalpayload 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 4

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

In [18]:
```
%sql select avg(payload_mass__kg_) as Average from SPACE
XTBL where booster_version like 'F9 v1.1%'
```

```
 * ibm_db_sa://mmp08973:***@54a2f15b-5c0f-46df-8954-7e38
e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:
32733/BLUDB
Done.
```

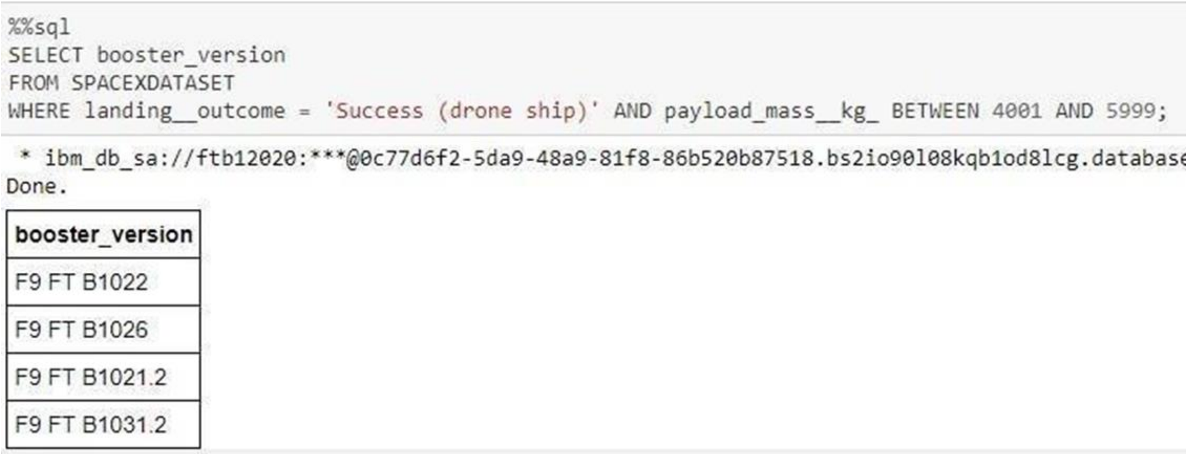
Out[18]:

average
3226

AVERAGE PAYLOAD MASS BY F9V1.1

- 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 endof our payload mass range

SUCCESSFULDRONE SHIP LANDING WITH PAYLOAD BETWEEN 4000 AND6000

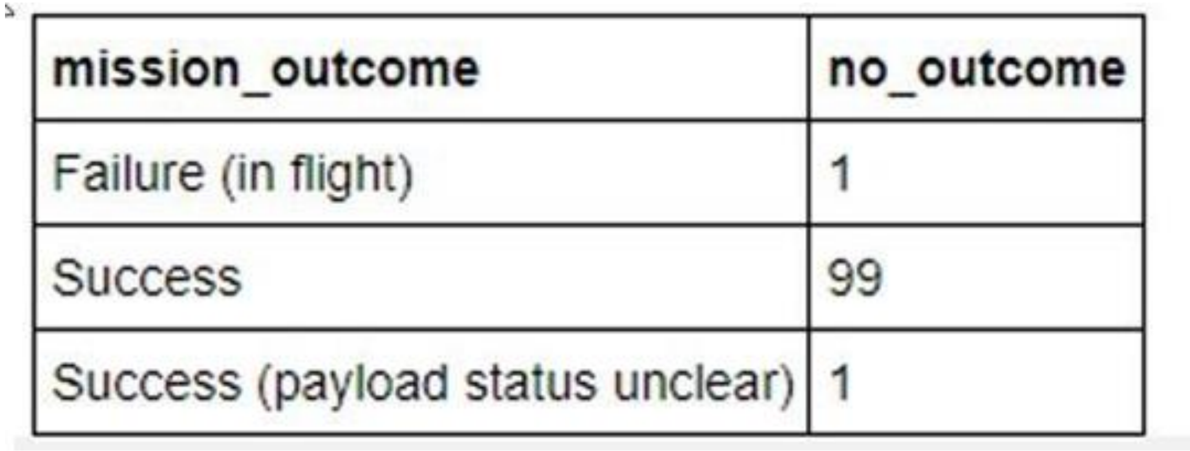


```
%%sql
SELECT booster_version
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4001 AND 5999;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.database
Done.
```

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

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

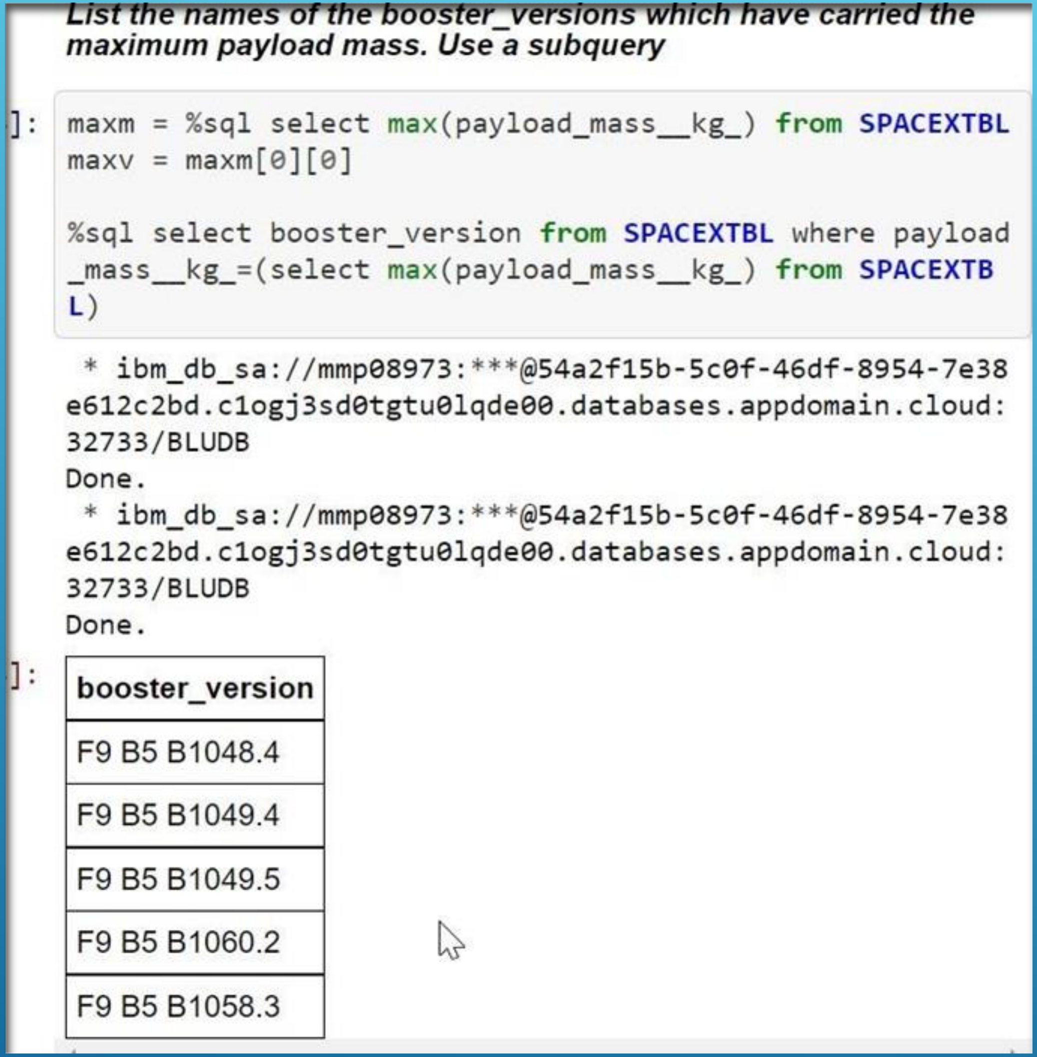


mission_outcome	no_outcome
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

TOTAL NUMBER OF EACHMISSIONOUTCOME

- This query returns a count ofeach
- mission outcome.
- SpaceX appears to achieve its mission outcome nearly 99% of the time.
- This means that most of the landing
- failures are intended.
- Interestingly, one launch has an unclear payload status and unfortunately one failed in flight.

BOOSTERS THAT CARRIED MAXIMUMPAYLOAD



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 SPACEXTB
   L)
```

```
 * ibm_db_sa://mmp08973:***@54a2f15b-5c0f-46df-8954-7e38
e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:
32733/BLUDB
Done.
 * ibm_db_sa://mmp08973:***@54a2f15b-5c0f-46df-8954-7e38
e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:
32733/BLUDB
Done.
```

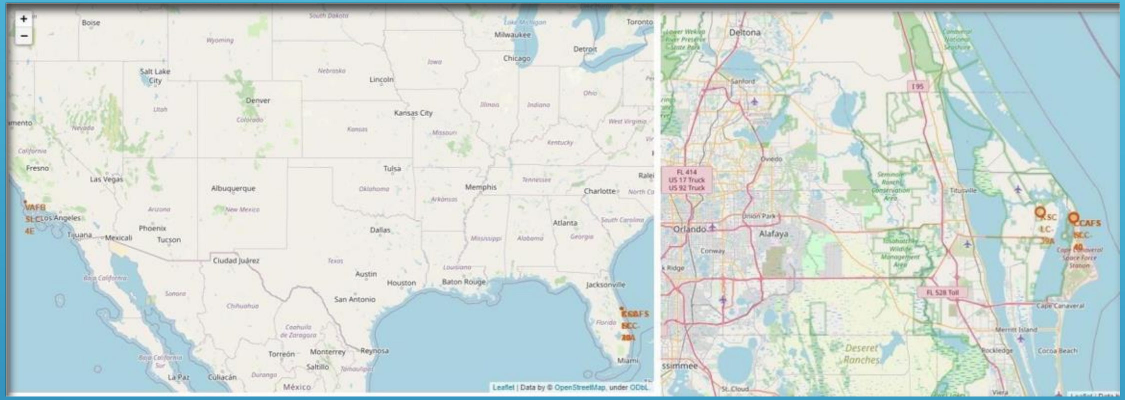
]:

booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3

This query returns the booster versions that carried the highest payload mass of 15600 kg.

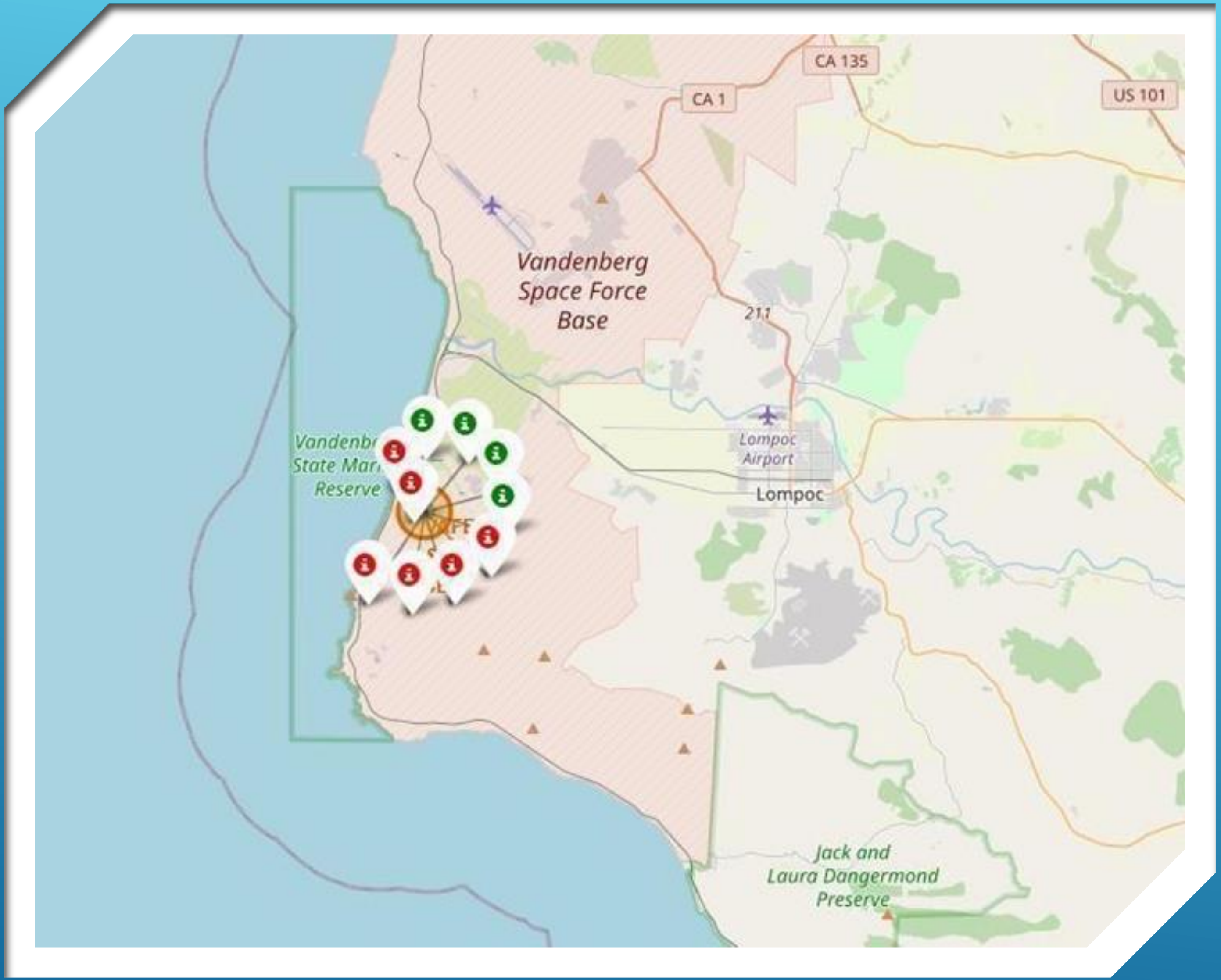
These booster versions are very similar and all are of the F9 B5 B10xx.xvariety.

This likely indicates payload mass correlates with the booster version that is used.



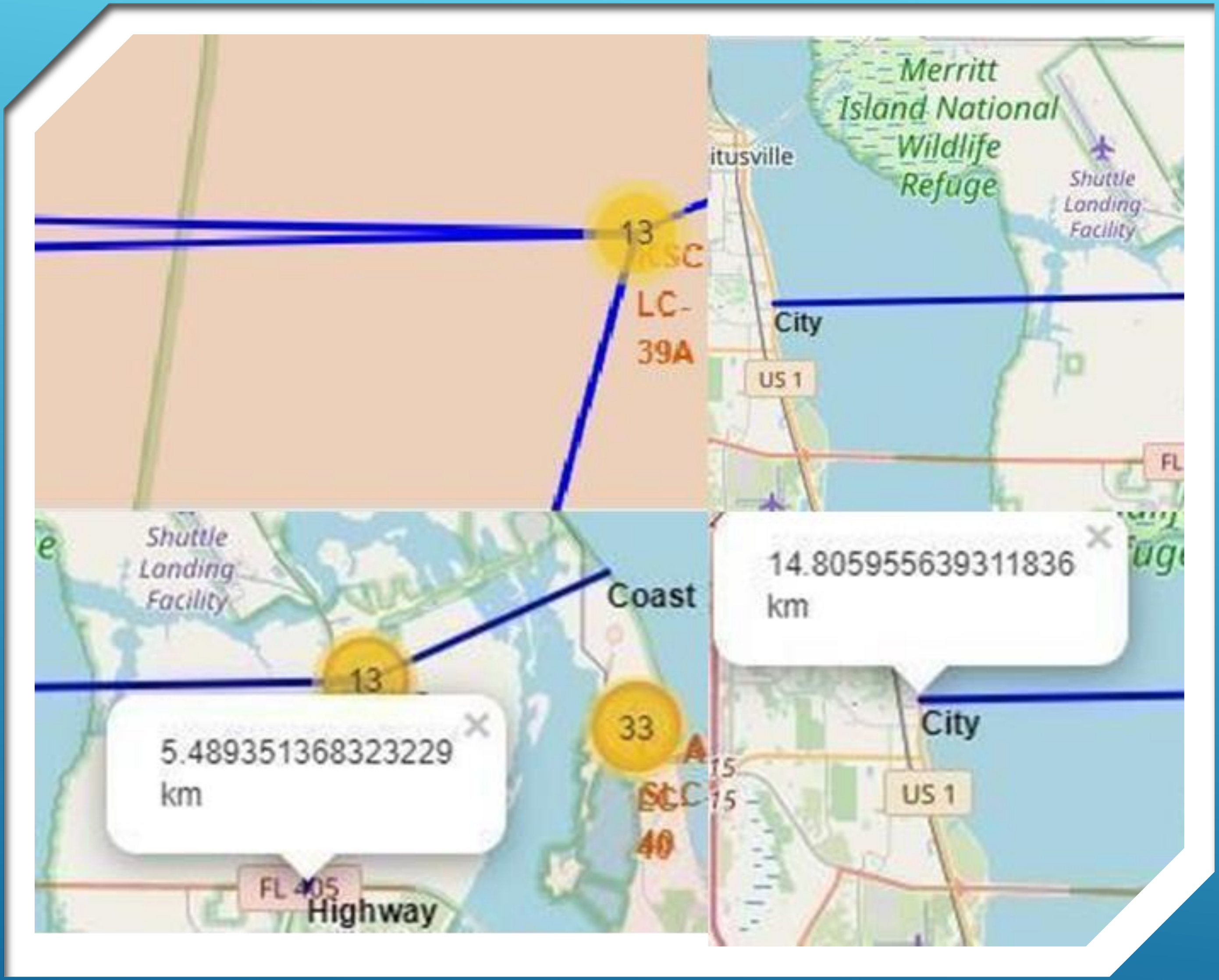
The left map shows all launch sites relative US map. The right map shows the two Florida launch sites since they are very close to each other. All launch sites are near the ocean.

LAUNCH SITE LOCATIONS


CA 135
CA 1
US 101
Vandenberg Space Force Base
211
Vandenbe State Mari Reserve
Lompoc Airport
Lompoc
Jack and Laura Dangermond Preserve

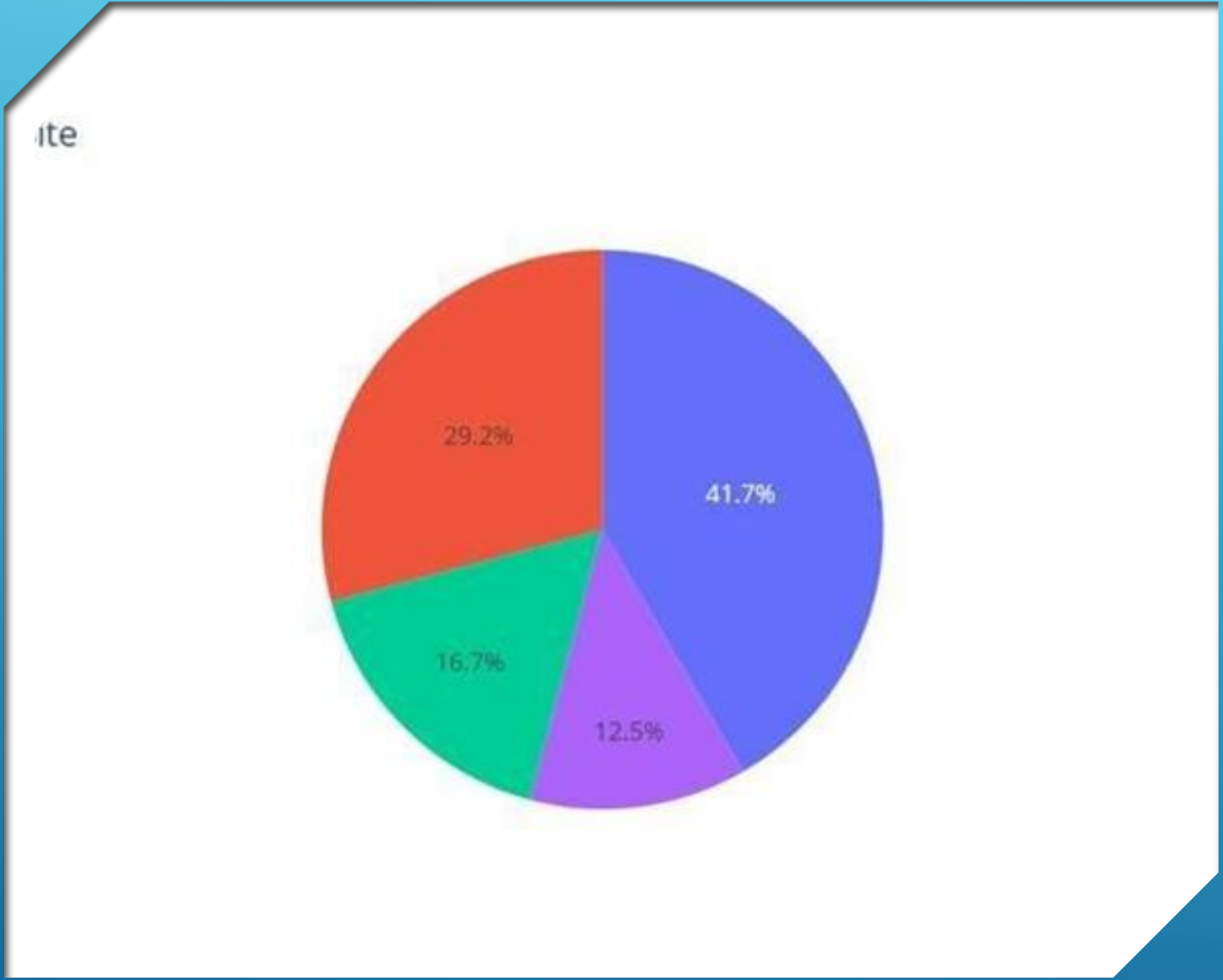
COLOR-CODED LAUNCH MARKERS

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.



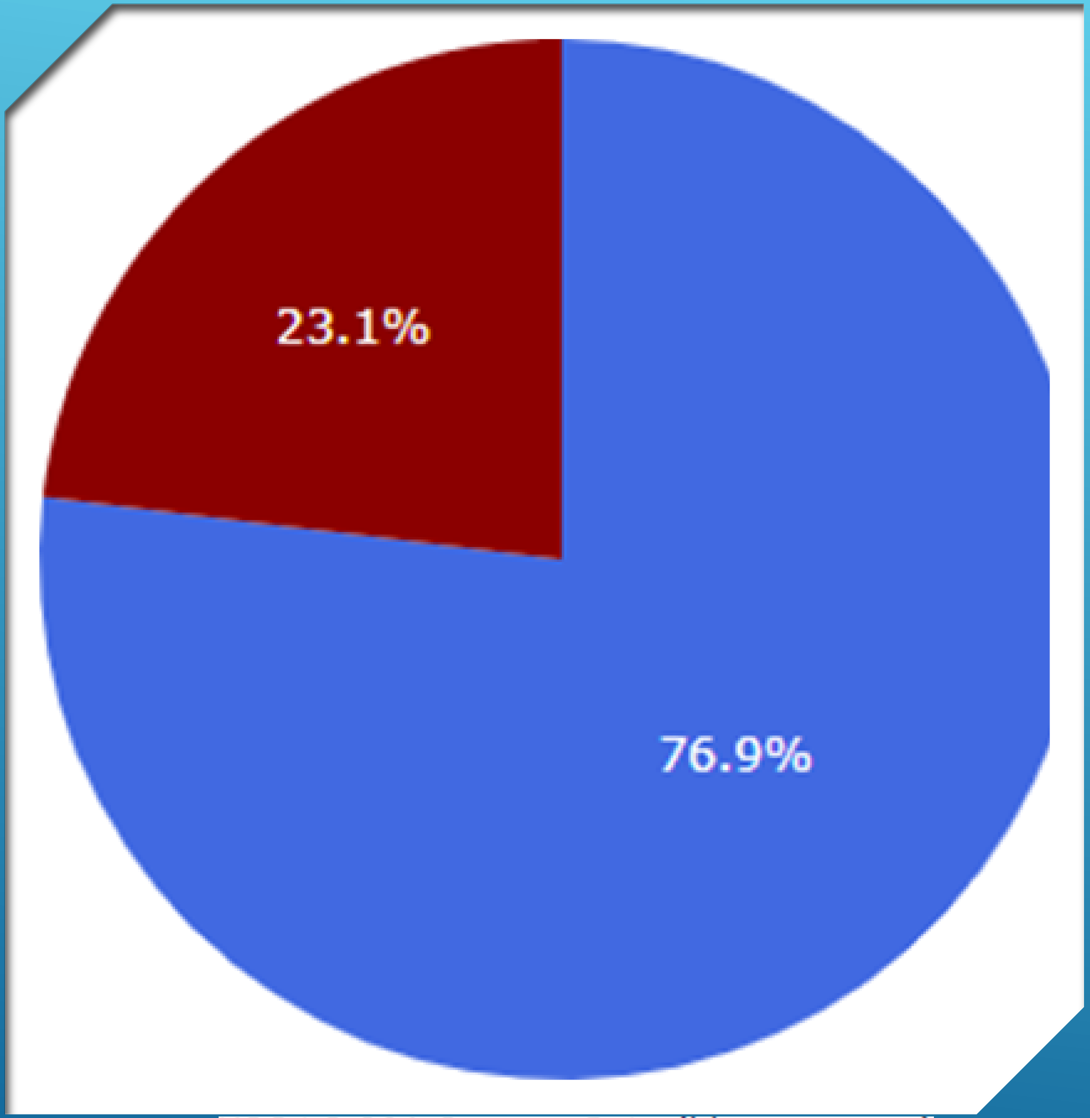
KEY LOCATION PROXIMITIES

Using KSC LC-39A as an example, launch sites are very close to railways for large part and supply transportation. Launch sites are close to highways for human and supply transport. Launch sites are also close to coasts and relatively far from cities so that launch failures can land in the sea to avoid rockets falling on densely populated areas.



SUCCESSFUL LAUNCHESACROSS LAUNCH SITES

This is the distribution of successful landings across all launch sites. CCAFS LC-40 is the old name of CCAFS SLC-40 so CCAFS and KSC have the same amount of successful landings, but a majority of the successful landings where performed before the name change. VAFB has the smallest share of successful landings. This may be due to smaller sample and increase in difficulty of launching in the westcoast.

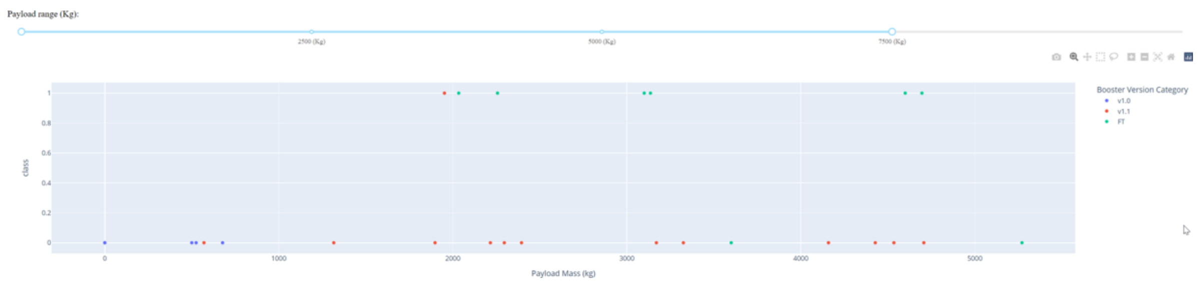


KSC LC-39A Success Rate (blue=success)

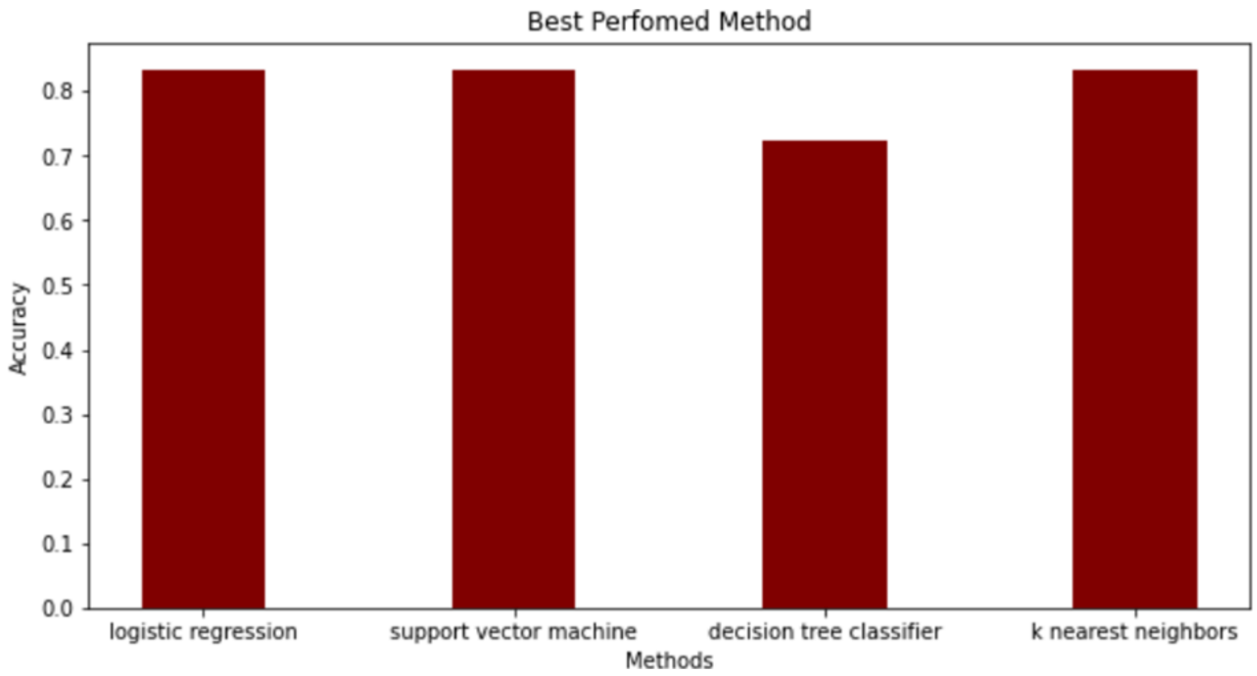
KSC LC-39A has the highest success rate with 10 successful landings and 3 failed landings.

HIGHEST SUCCESSRATELAUNCHSITE

PAYLOAD MASS VS. SUCCESSVS. BOOSTER VERSION CATEGORY



Plotly dashboard has a Payload range selector. However, this is set from 0-10000 instead of the max Payload of 15600. Class indicates 1 for successful landing and 0 for failure. Scatter plot also accounts for booster version category in color and number of launches in point size. In this particular range of 0-7500, interestinglythere are two failed landings with payloads of zero kg.

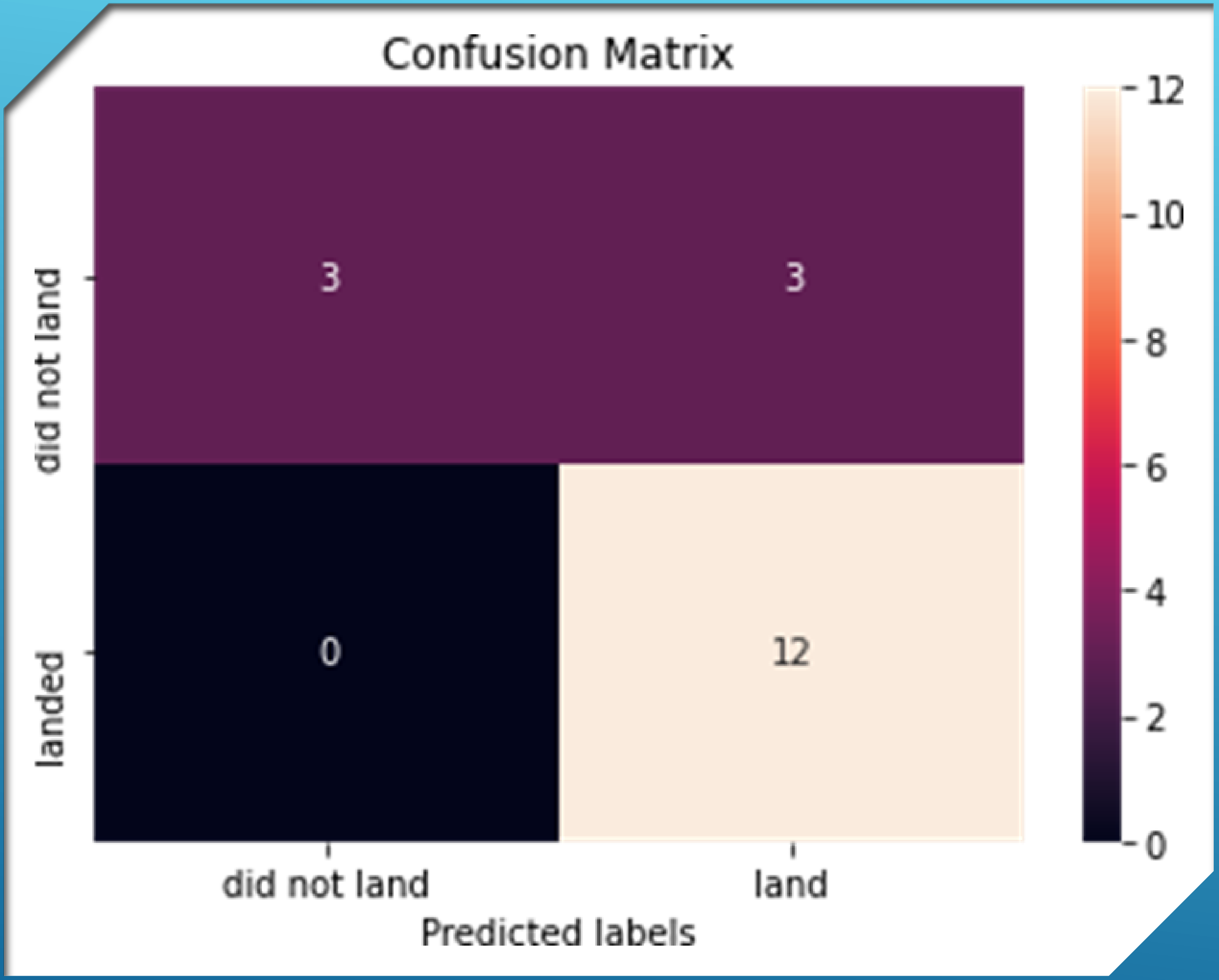


CLASSIFICATIONACCURACY

The models had virtually the same accuracy on the test set at 83.33%accuracy, except the decision tree classifier with 72,23%.

It should be noted that test size is small at only sample size of 18.

This can cause large variance in accuracy results, such as those in Decision TreeClassifier model in repeated runs.



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).

CONCLUSIONS

The task was to develop a machine learning model for SpaceX . The goal of the model is to predict when Stage 1 will successfully land to save ~$100M USD. Using data from a public SpaceX API and web scraping the SpaceX Wikipedia page.Creating a dashboard for visualization.Creating data labels and storing the data in a DB2 SQL database. The best launch site is the KSC LC-39A launches weighing more than 7,000 kg and are less risky. Although most missions are successful, the results of successful landings appear to improve over time as processes and rockets evolve.Decision tree classifier can be used to predict successful landings and increase profits. SpaceX can use this model to predict with relatively high accuracy whether a launch will have a successful Stage 1 landing before the launch occurs to determine whether or not the launch should proceed. More data should be collected to determine the best machine learning model and improve accuracy

APPENDIX

https://github.com/Cem0061/Applied-Data-Science-Capstone

Special Thanks to All Instructors