IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jimmy Zhang

June 3, 2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results (EDA, Proximity analysis, Dashboard, Predictive modeling)
- Conclusion
- Appendix

Executive Summary

- Background and Objectives:
  - Reusing first stage of rockets optimal in reducing costs of space launch
  - Determining vital parameters to successfully recover first stage after use
- Methodology:
  - Obtained launch data from SpaceX and Wikipedia
  - Conducted EDA with Python (visualization), SQL, Folium
  - Presented data with Dash, created and tested classification model
- Results:
  - Different orbit types, booster versions suitable for different ranges of payload mass
  - Location landmarks (coastline, city) may be important to success
  - Decision Tree model yields best classification

Introduction

- In advent of space age, minimizing costs of space launch is vital
- SpaceX model of reusing first stage of rocket critical to low-cost approach

- **Problem: How to optimize landing and recovery of first stage after use**
- Investigation of SpaceX historical data to understand potential parameters integral to successful first stage landing

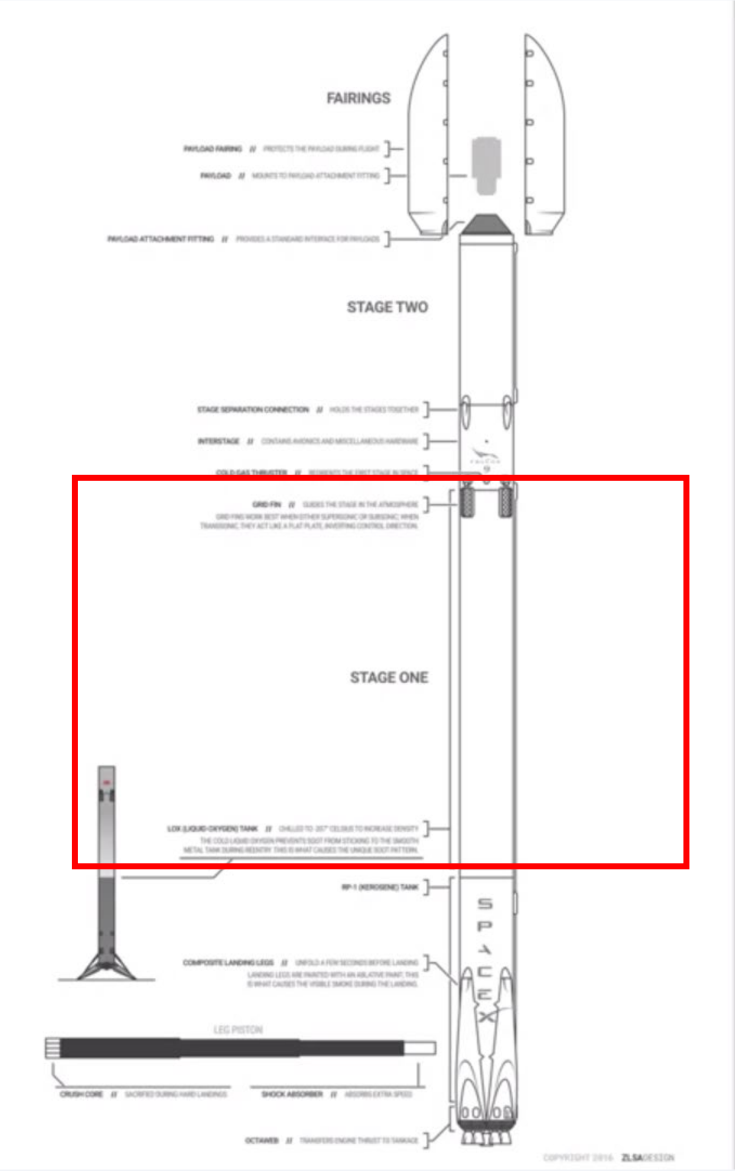

FAIRINGS
PAYLOAD FAIRING // PROTECTS THE PAYLOAD DURING FLIGHT
PAYLOAD // MOUNTS TO PAYLOAD ATTACHMENT FITTING
PAYLOAD ATTACHMENT FITTING // PROVIDES A STANDARD INTERFACE FOR PAYLOADS
STAGE TWO
STAGE SEPARATION CONNECTION // HOLDS THE STAGES TOGETHER
INTERSTAGE // CONTAINS AVIONICS AND MISCELLANEOUS HARDWARE
COLD GAS THRUSTER // REORIENTS THE FIRST STAGE IN SPACE
GRID FIN // GUIDES THE STAGE IN THE ATMOSPHERE
GRID FINS WORK BEST WHEN EITHER SUPERSONIC OR SUBSONIC; WHEN TRANSONIC, THEY ACT LIKE A FLAT PLATE, INVERTING CONTROL DIRECTION.
STAGE ONE
LOX (LIQUID-OXYGEN) TANK // CHILLED TO -207° CELSIUS TO INCREASE DENSITY
THE COLD LIQUID OXYGEN PREVENTS SOOT FROM STICKING TO THE SMOOTH METAL TANK DURING REENTRY. THIS IS WHAT CAUSES THE UNIQUE SOOT PATTERN.
RP-1 (KEROSENE) TANK
SPACEX
COMPOSITE LANDING LEGS // UNFOLD A FEW SECONDS BEFORE LANDING
LANDING LEGS ARE PAINTED WITH AN ABLATIVE PAINT. THIS IS WHAT CAUSES THE VISIBLE SMOKE DURING THE LANDING.
LEG PISTON
CRUSH CORE // SACRIFICED DURING HARD LANDINGS
SHOCK ABSORBER // ABSORBS EXTRA SPEED
OCTAWEB // TRANSFERS ENGINE THRUST TO TANKAGE
COPYRIGHT 2016 ZLSADESIGN

Image from: Forest Katsch

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Obtained from list of launch records of Falcon 9 from Wikipedia via web scraping (Python) and from SpaceX API via requests from Python
- Perform data wrangling
  - Convert HTML tables to usable data frame
  - Replaced missing values
  - Conducted one-hot encoding for launch success, type of orbit, launch site, landing pad, and serial code
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Standardized data, split data into training/test data, and perform GridSearchCV to optimize models (logistic regression, support vector machine, decision tree, K-nearest neighbor)

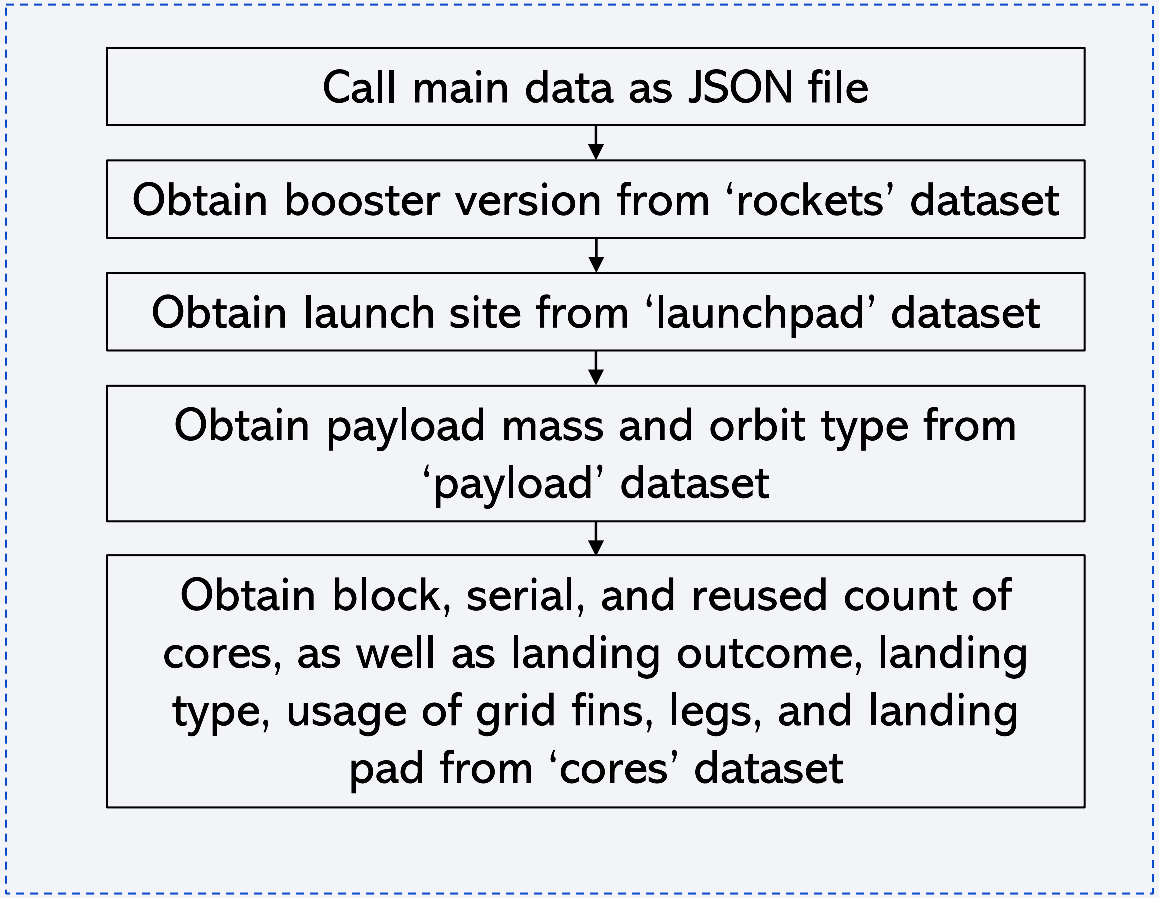
Data Collection

Two sets of data were collected from:

- SpaceX API (api.spacexdata.com), containing:
  - Flight Number, Launch Date, Booster Version, Payload Mass, Orbit Type
  - Launch Site, Launch Outcome, Landing Pad, Longitude, Latitude
  - Flights Flown, Grid Fins Used, Core (Reused Count, Block, Serial), Legs Used
- Wikipedia (https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches), adding:
  - Payload Type
  - Customer
  - Booster Landing Outcome

Data Collection – SpaceX API

- Collected data from multiple datasets in API (rockets, launchpad, payload, and core)
- GitHub URL: https://github.com/jimmyz100/SpaceXDataScience/blob/6d94a7cac1dbe5928f71e9ba8bef841f8a95a74d/Data%20Collection%20API%20.ipynb



Call main data as JSON file

Obtain booster version from ‘rockets’ dataset

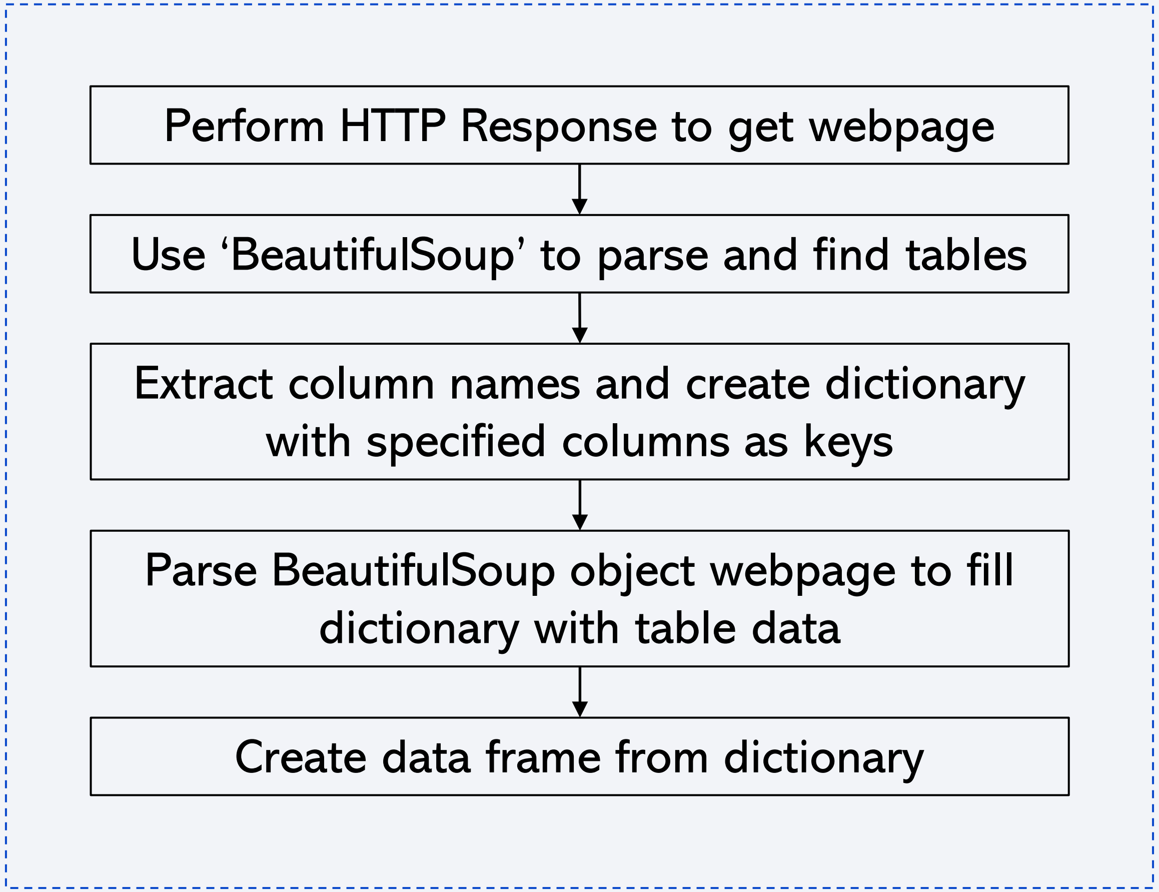
Obtain launch site from ‘launchpad’ dataset

Obtain payload mass and orbit type from ‘payload’ dataset

Obtain block, serial, and reused count of cores, as well as landing outcome, landing type, usage of grid fins, legs, and landing pad from ‘cores’ dataset

Data Collection – Wikipedia Scraping

- Obtained flight no., launch site, payload type, payload mass, orbit, customer, launch outcome, booster version, booster landing outcome, launch date and time
- GitHub URL: https://github.com/jimmyz100/SpaceXDataScience/blob/63831b5a5fd70e44099c35287de66a463c126fee/Data_Collection_with_Web_Scraping.ipynb



Perform HTTP Response to get webpage

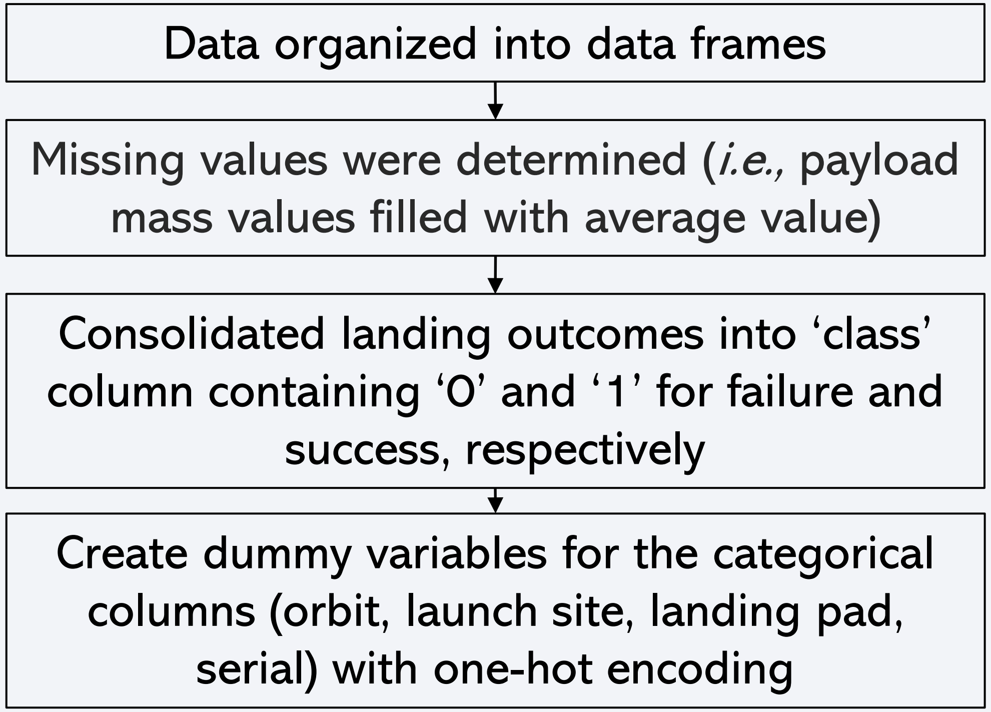
Use ‘BeautifulSoup’ to parse and find tables

Extract column names and create dictionary with specified columns as keys

Parse BeautifulSoup object webpage to fill dictionary with table data

Create data frame from dictionary

Data Wrangling



Data organized into data frames

Missing values were determined (*i.e.,* payload mass values filled with average value)

Consolidated landing outcomes into ‘class’ column containing ‘0’ and ‘1’ for failure and success, respectively

Create dummy variables for the categorical columns (orbit, launch site, landing pad, serial) with one-hot encoding

GitHub URLs:
https://github.com/jimmyz100/SpaceXDataScience/blob/e9e8301744ce790fabdf7ae22be7845ad6cad36e/Data_Collection_with_Web_Scraping.ipynb,
https://github.com/jimmyz100/SpaceXDataScience/blob/e9e8301744ce790fabdf7ae22be7845ad6cad36e/labs_jupyter_spacex_Data_wrangling.ipynb,
https://github.com/jimmyz100/SpaceXDataScience/blob/e9e8301744ce790fabdf7ae22be7845ad6cad36e/jupyter_labs_eda_dataviz.ipynb

EDA with Data Visualization

- Charts plotted:
  - Payload mass vs Flight number: Understanding progression of payload size over time and contribution to success
  - Launch site vs Flight number: Understanding launch site usage over time and contribution to success
  - Launch site vs Payload mass: Understanding launch site preference for payload size and contribution to success
  - Success rate for orbit type: Understanding which orbit type yields highest success
  - Flight number/payload mass vs Orbit type: Understanding how orbit type correlates over time and with payload size in successful flights
  - Yearly success rate: Determine success over time of program

GitHub URL:
https://github.com/jimmyz100/SpaceXDataScience/blob/e9e8301744ce790fabdf7ae22be7845ad6cad36e/jupyter_labs_eda_dataviz.ipynb

EDA with SQL

- SQL queries include:
  - Determining all launch site names
  - Payload mass based on booster/customer
  - Landing outcome type (ocean, drone ship, ground pad, parachute) and count
  - Time of successful launches

GitHub URL:
https://github.com/jimmyz100/SpaceXDataScience/blob/e5f06de045dec08338952a62bf237b4cdea7ae2b/jupyter_labs_eda_sql_coursera.ipynb

Build an Interactive Map with Folium

- Map objects added:
  - Circles and markers to indicate launch site locations
  - Markers to indicate all launches at respective locations (green = success, red = failure): To highlight success rate and number of launches per site
  - Markers to indicate closest coastline, railway, highway, and city to each site, lines drawn to indicate distance: To determine how close these landmarks are to each site

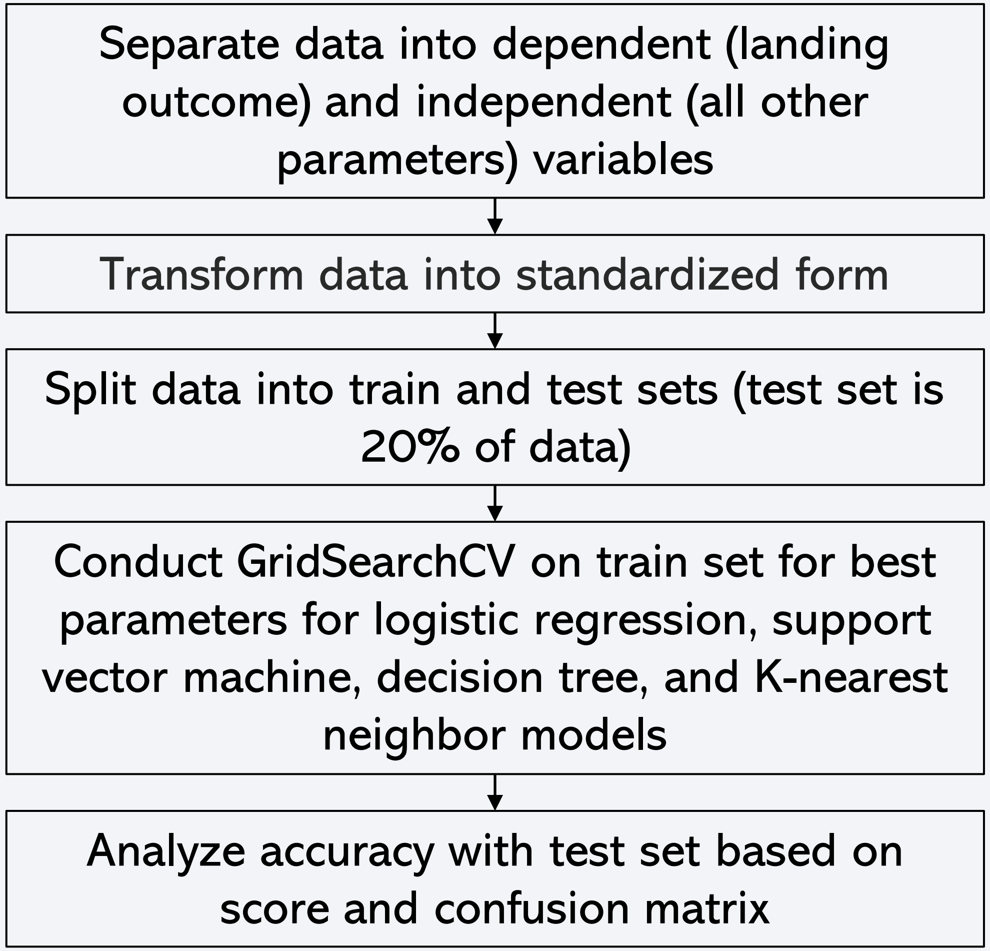
GitHub URL:
https://github.com/jimmyz100/SpaceXDataScience/blob/f5c084e08897cbd485ded4b5dd601967cd5296ca/lab_jupyter_launch_site_location%20(1).ipynb

Build a Dashboard with Plotly Dash

- Interactive Dashboard:
  - Pie chart of total success launches by launch site, as well as success/failure rate of each site (by selecting specific site)
  - Scatter plot of launch success by payload mass, booster version, and specified site (if applicable) – controlled by payload mass slider and site selection
- Provides interaction and analysis into the success rate and distribution of each site, payload range most inducive for success based on booster version and site

GitHub URL to Python Code:
https://github.com/jimmyz100/SpaceXDataScience/blob/cff9c7bc32b19efa160af7150dc4240ab5f9ac14/spacex_dash_app%20Final.py

Predictive Analysis (Classification)



Separate data into dependent (landing outcome) and independent (all other parameters) variables

Transform data into standardized form

Split data into train and test sets (test set is 20% of data)

Conduct GridSearchCV on train set for best parameters for logistic regression, support vector machine, decision tree, and K-nearest neighbor models

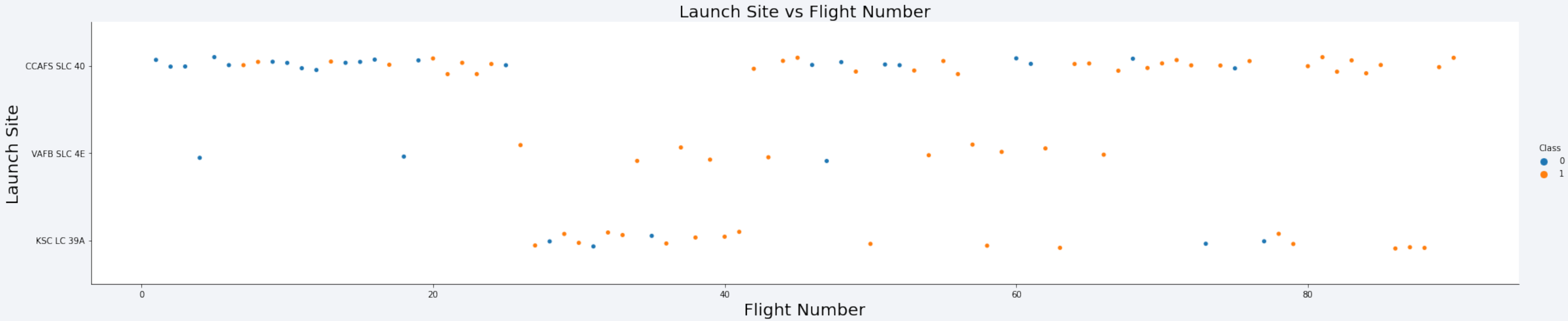
Analyze accuracy with test set based on score and confusion matrix

GitHub URL: https://github.com/jimmyz100/SpaceXDataScience/blob/18a9d117224321445adb66f02705303b9e9572a7/SpaceX_Machine_Learning_Prediction_Part_5.ipynb

Section 2

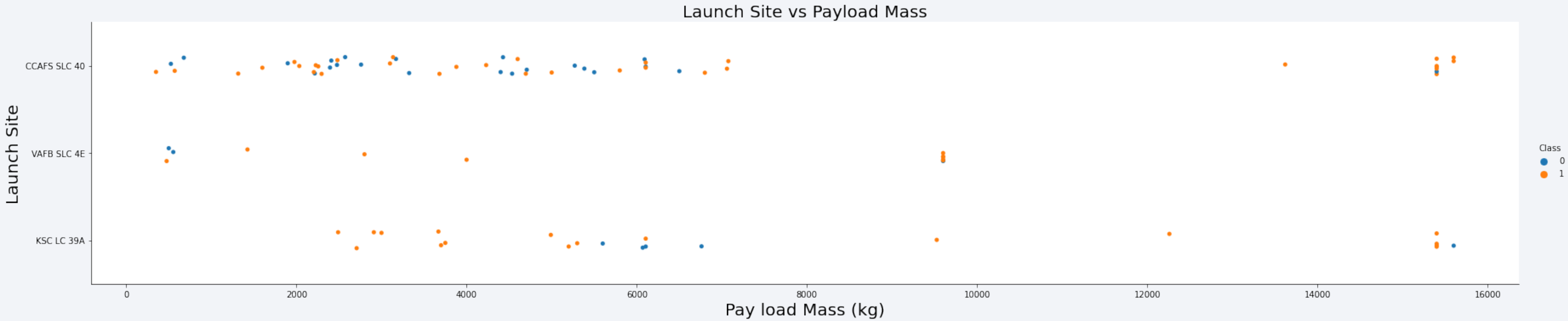
Insights drawn from EDA

Flight Number vs. Launch Site



- Most early (low flight number) and late (high flight number) flights located in CCAFS
- KSC flights tend to occur in later flights (high flight number)
- Each site yielded successful flights in later flights (at least 5 most recent flights from each site was successful)

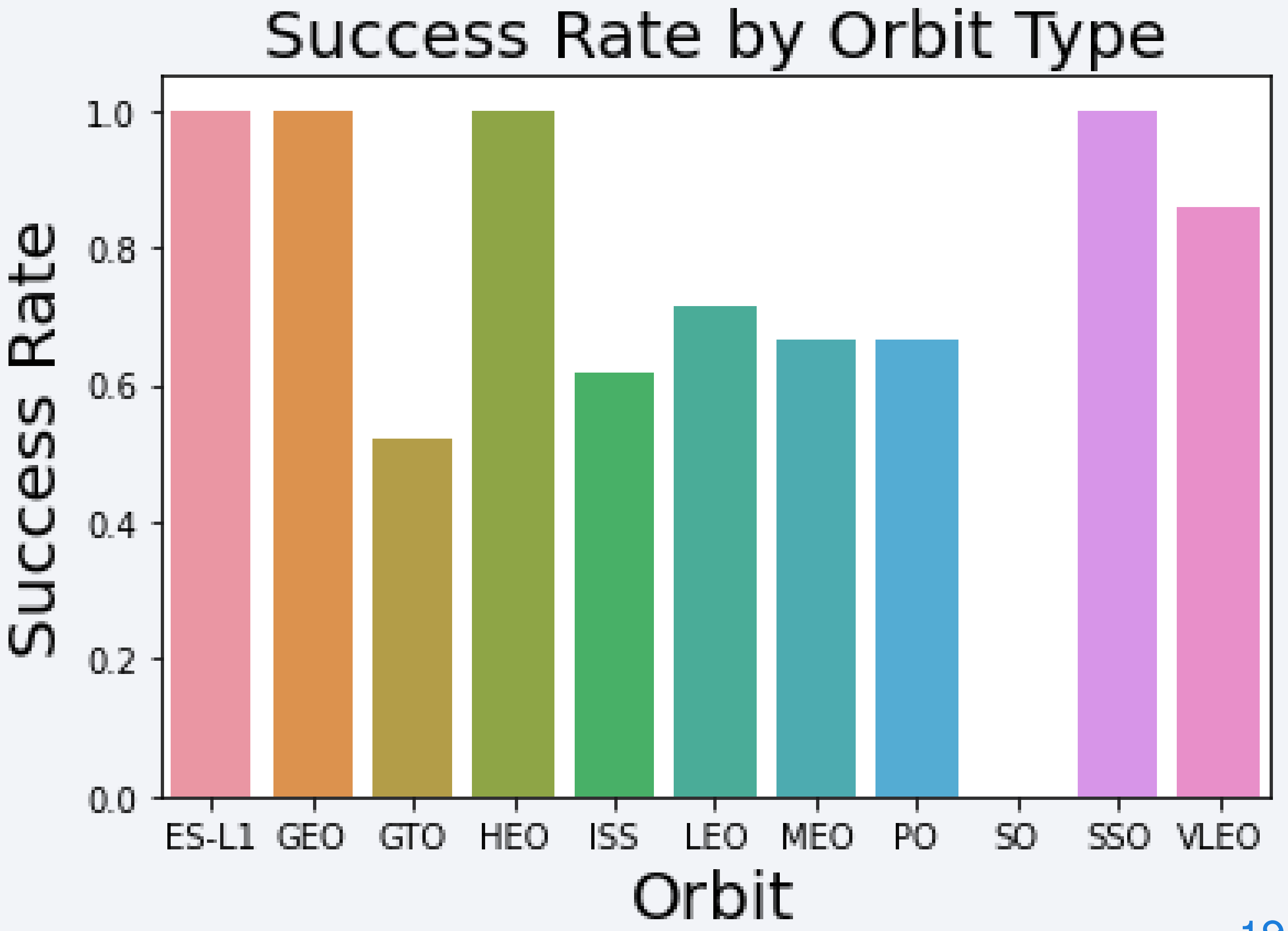
Payload vs. Launch Site



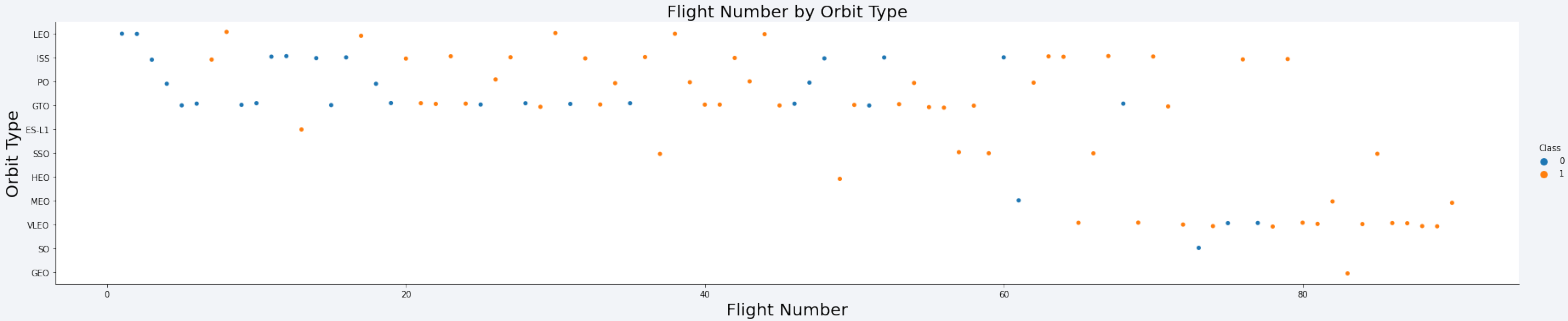
- Most flights from CCAFS are below 7000 kg., accounting for most flights in that payload range
- VAFB has no flights above 10000 kg. payload
- Higher payload mass (>10000 kg.) generally yielded higher success rate

Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, and SSO yielded 100% success rate
- GTO yielded 51.8% success (2nd lowest) – intermediate orbit to GEO
- Lower orbit (LEO, MEO) tend to yield lower success rate than GEO (higher orbit)
  - LEO- 71.4%
  - MEO- 66.7%

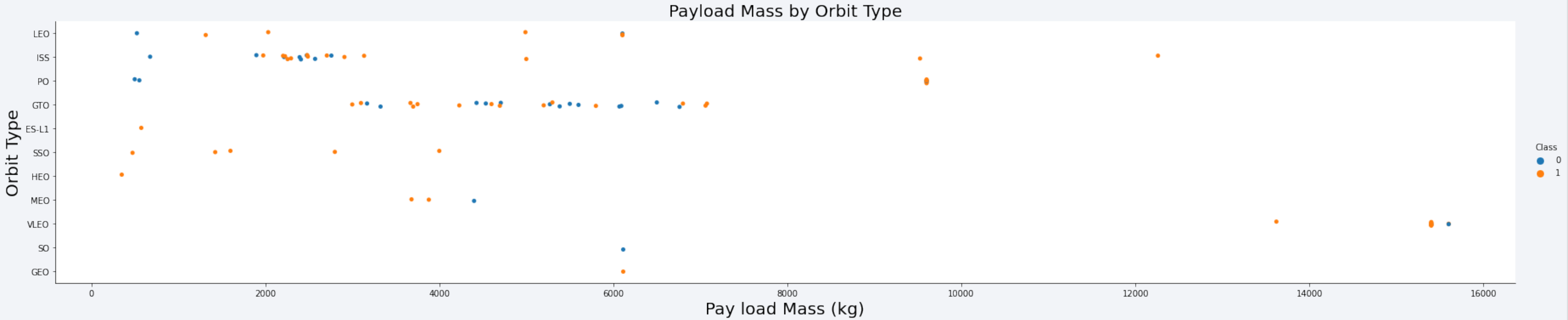


Flight Number vs. Orbit Type



- LEO occurs at earlier flights; HEO, VLEO, SO, and GEO occur at later flights
- For more frequent orbit types (LEO, GTO, VLEO, ISS), flights are more likely to be successful as flight number increases

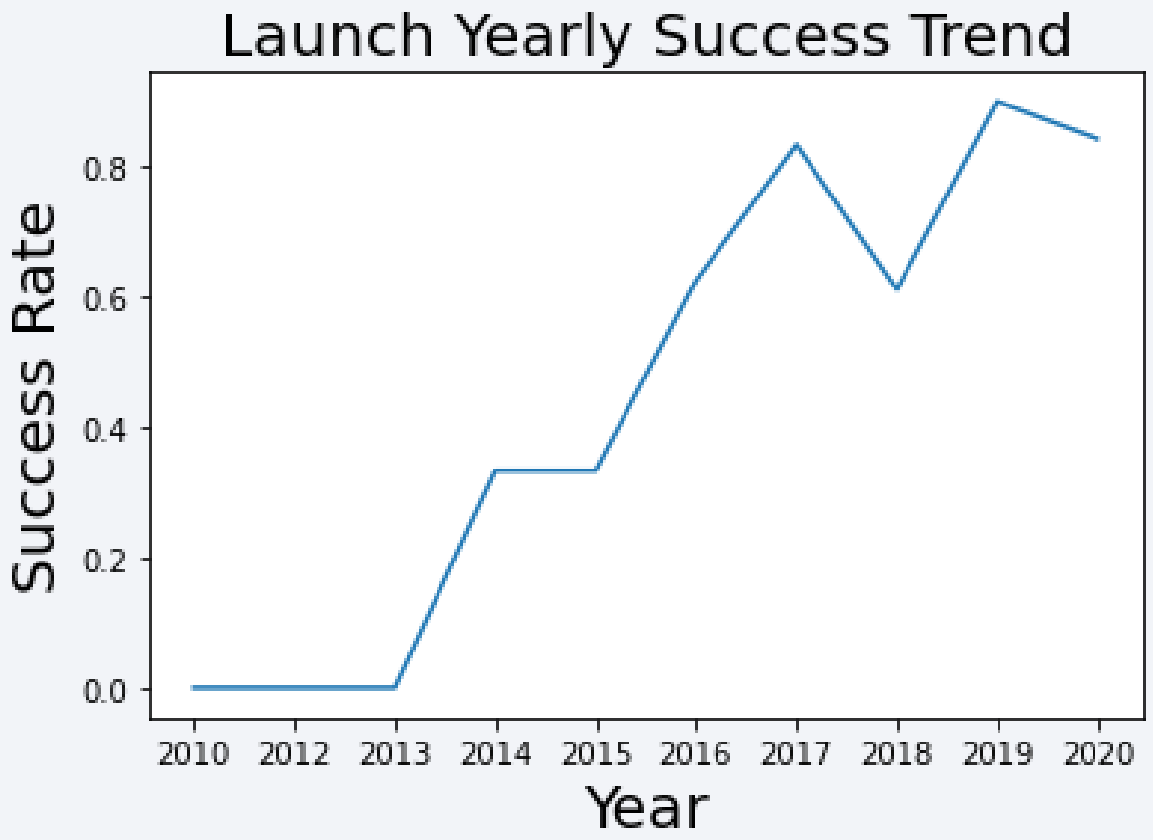
Payload vs. Orbit Type



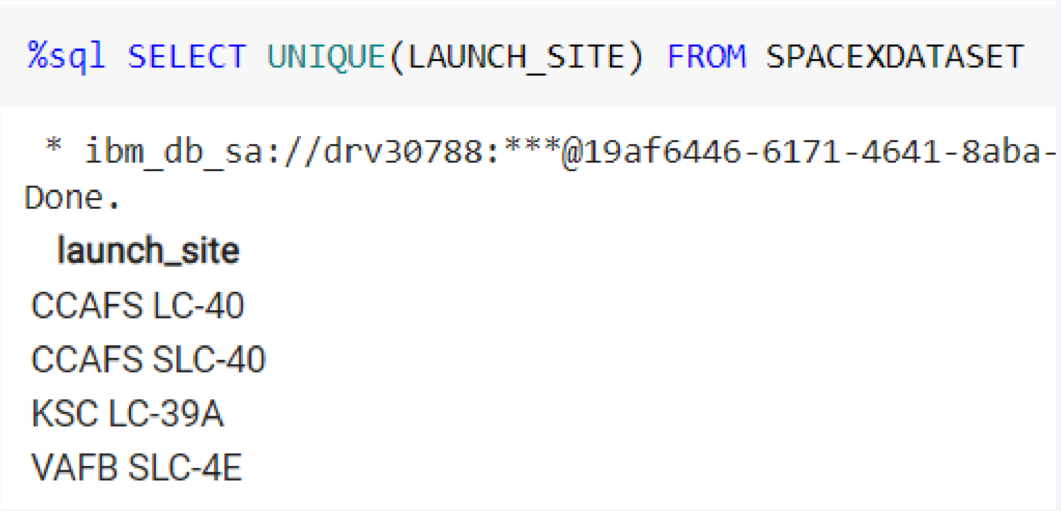
- LEO, GTO, ES-L1, SSO, HEO, MEO, and GEO carry low payload mass (<8000 kg.)
- VLEO carries mass >13000 kg.
- For LEO, ISS, PO, heavier payload mass tend to yield more successful launches

Launch Success Yearly Trend

- Success rate has trended higher over time, with the exceptions of 2018 and 2020
- Peak success rate in 2019



All Launch Site Names



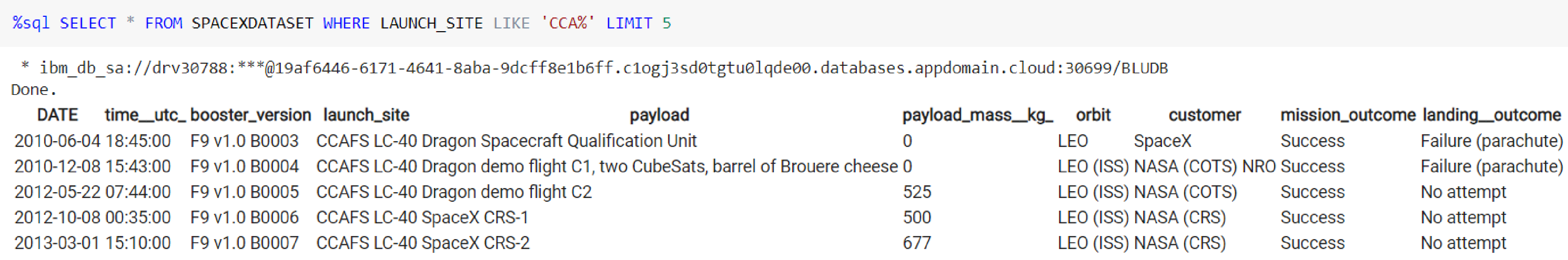
```
%sql SELECT UNIQUE(LAUNCH_SITE) FROM SPACEXDATASET
```

```
 * ibm_db_sa://drv30788:***@19af6446-6171-4641-8aba-
Done.
  launch_site
 CCAFS LC-40
 CCAFS SLC-40
 KSC LC-39A
 VAFB SLC-4E
```

Launch Sites:

- CCAFS LC-40/SLC-40 – Cape Canaveral Air Force Station (Florida)
- KSC LC-39A – Kennedy Space Center (Florida)
- VAFB SLC-4E – Vandenberg Space Force Base (California)

Launch Site Names Begin with 'CCA'



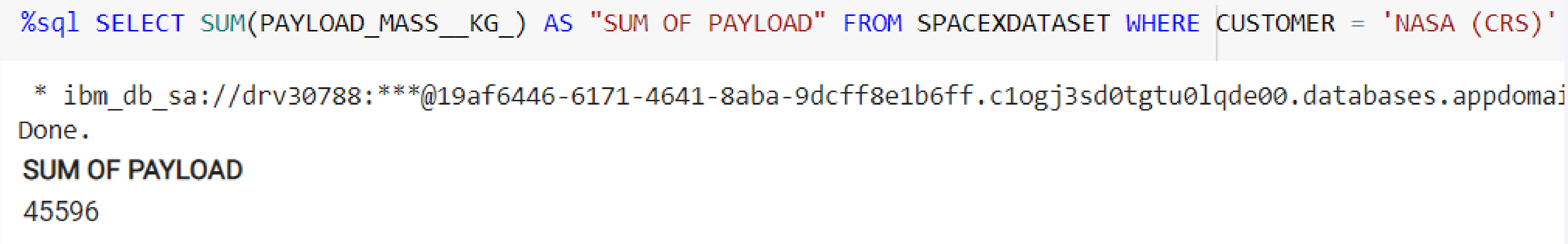
```
%sql SELECT * FROM SPACEXDATASET WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
```

```
 * ibm_db_sa://drv30788:***@19af6446-6171-4641-8aba-9dcff8e1b6ff.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30699/BLUDB
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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

- Launch site names that begin with ‘CCA’ correspond to the CCAFS launch site
- SLC-40 and LC-40 refer to same launch site (change in nomenclature only)

Total Payload Mass from NASA (CRS) Boosters

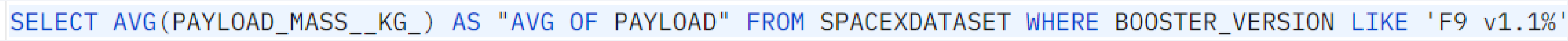


```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS "SUM OF PAYLOAD" FROM SPACEXDATASET WHERE CUSTOMER = 'NASA (CRS)'
```

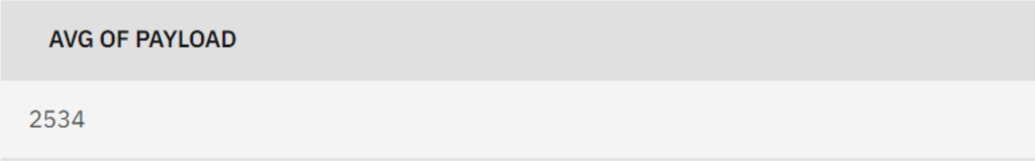
```
 * ibm_db_sa://drv30788:***@19af6446-6171-4641-8aba-9dcff8e1b6ff.c1ogj3sd0tgtu0lqde00.databases.appdomai
Done.
SUM OF PAYLOAD
45596
```

- 44596 kg., ranked third largest aggregate payload mass, behind SpaceX and Iridium Communications
- Highest aggregate payload mass of all NASA divisions

Average Payload Mass by F9 v1.1



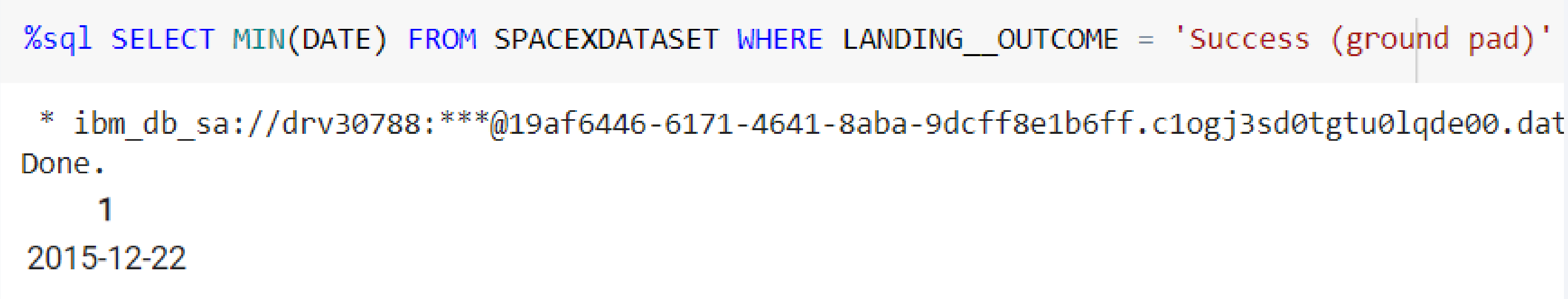
```
SELECT AVG(PAYLOAD_MASS__KG_) AS "AVG OF PAYLOAD" FROM SPACEXDATASET WHERE BOOSTER_VERSION LIKE 'F9 v1.1%'
```



AVG OF PAYLOAD
2534

- Average payload of 2534 kg., second least among all booster versions

First Successful Ground Landing Date

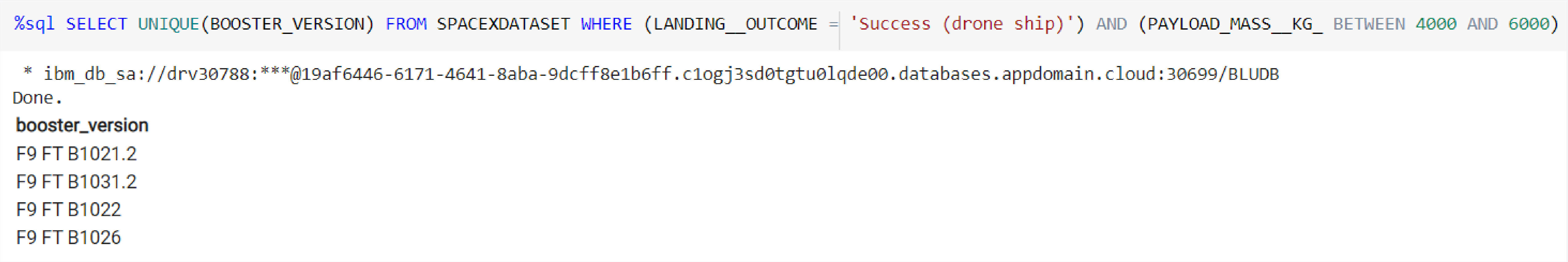


```
%sql SELECT MIN(DATE) FROM SPACEXDATASET WHERE LANDING__OUTCOME = 'Success (ground pad)'
```

```
 * ibm_db_sa://drv30788:***@19af6446-6171-4641-8aba-9dcff8e1b6ff.c1ogj3sd0tgtu0lqde00.dat
Done.
    1
2015-12-22
```

- December 12, 2015 – first successful ground pad landing
- 5 years after first Falcon 9 launch
- Occurring after successful ocean landing, before successful drone ship landing

Successful Drone Ship Landing with Payload between 4000 and 6000



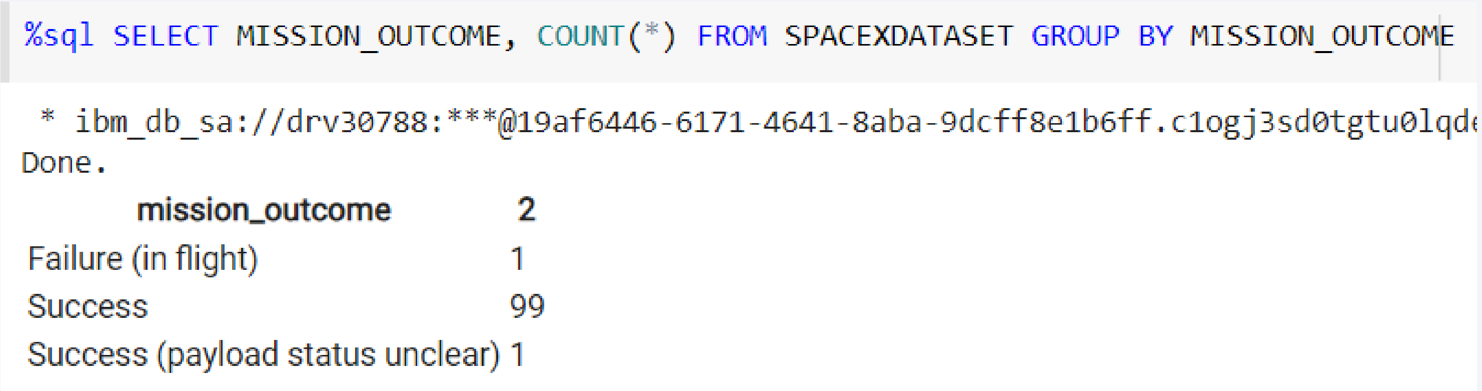
```
%sql SELECT UNIQUE(BOOSTER_VERSION) FROM SPACEXDATASET WHERE (LANDING__OUTCOME = 'Success (drone ship)') AND (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000)
```

```
 * ibm_db_sa://drv30788:***@19af6446-6171-4641-8aba-9dcff8e1b6ff.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30699/BLUDB
Done.
```

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

- All boosters are F9 FT
- Version B1021 (2016) notable for first successful drone ship landing

Total Number of Successful and Failure Mission Outcomes



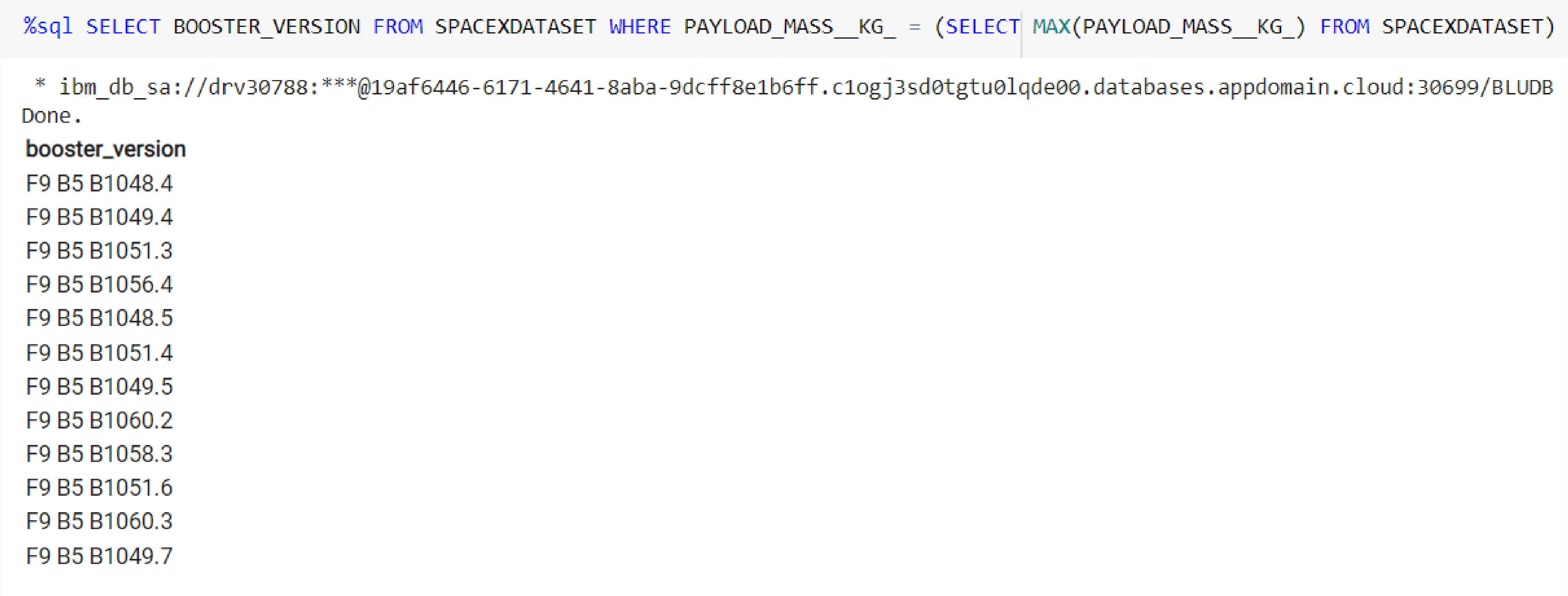
```
%sql SELECT MISSION_OUTCOME, COUNT(*) FROM SPACEXDATASET GROUP BY MISSION_OUTCOME
```

```
 * ibm_db_sa://drv30788:***@19af6446-6171-4641-8aba-9dcff8e1b6ff.c1ogj3sd0tgtu0lqde
Done.
```

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- Almost all missions were successful (99 out of 101)
    - Only one successful mission yielded unsuccessful payload deployment (Zuma mission)
- Only failure resulted from over pressurization due to faulty support hardware

Boosters Carried Maximum Payload

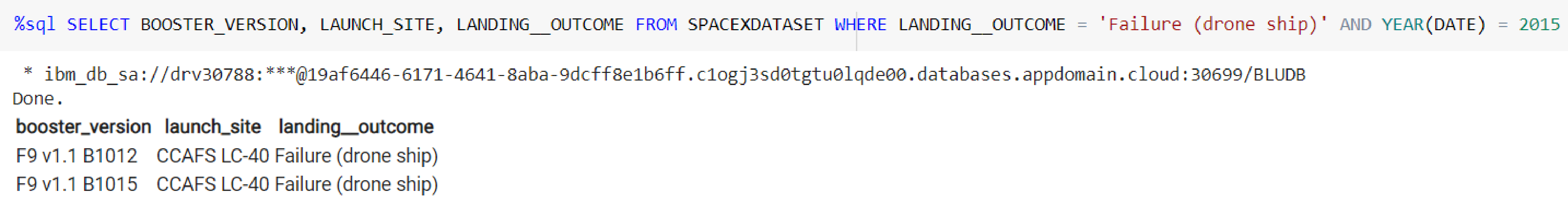


```
%sql SELECT BOOSTER_VERSION FROM SPACEXDATASET WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXDATASET)
```

```
 * ibm_db_sa://drv30788:***@19af6446-6171-4641-8aba-9dcff8e1b6ff.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30699/BLUDB
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
```

- All boosters are F9 B5 (latest version in dataset)
- Max payload of 15600 kg.

2015 Launch Records for Landing Failures



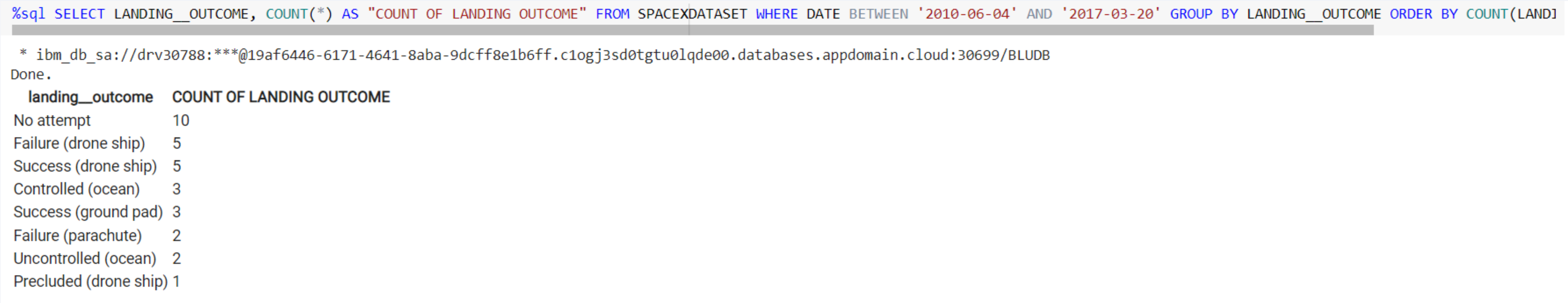
```
%sql SELECT BOOSTER_VERSION, LAUNCH_SITE, LANDING__OUTCOME FROM SPACEXDATASET WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND YEAR(DATE) = 2015
```

```
 * ibm_db_sa://drv30788:***@19af6446-6171-4641-8aba-9dcff8e1b6ff.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30699/BLUDB
Done.
```

booster_version	launch_site	landing__outcome
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- Outcome correspond to F9 v1.1 booster version, flown from CCAFS LC-40
- F9 v1.1 most likely due to precursor to F9 FT, first to successfully land on drone ship

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



```
%sql SELECT LANDING__OUTCOME, COUNT(*) AS "COUNT OF LANDING OUTCOME" FROM SPACEXDATASET WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING__OUTCOME ORDER BY COUNT(LANDI
```

```
 * ibm_db_sa://drv30788:***@19af6446-6171-4641-8aba-9dcff8e1b6ff.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30699/BLUDB
Done.
```

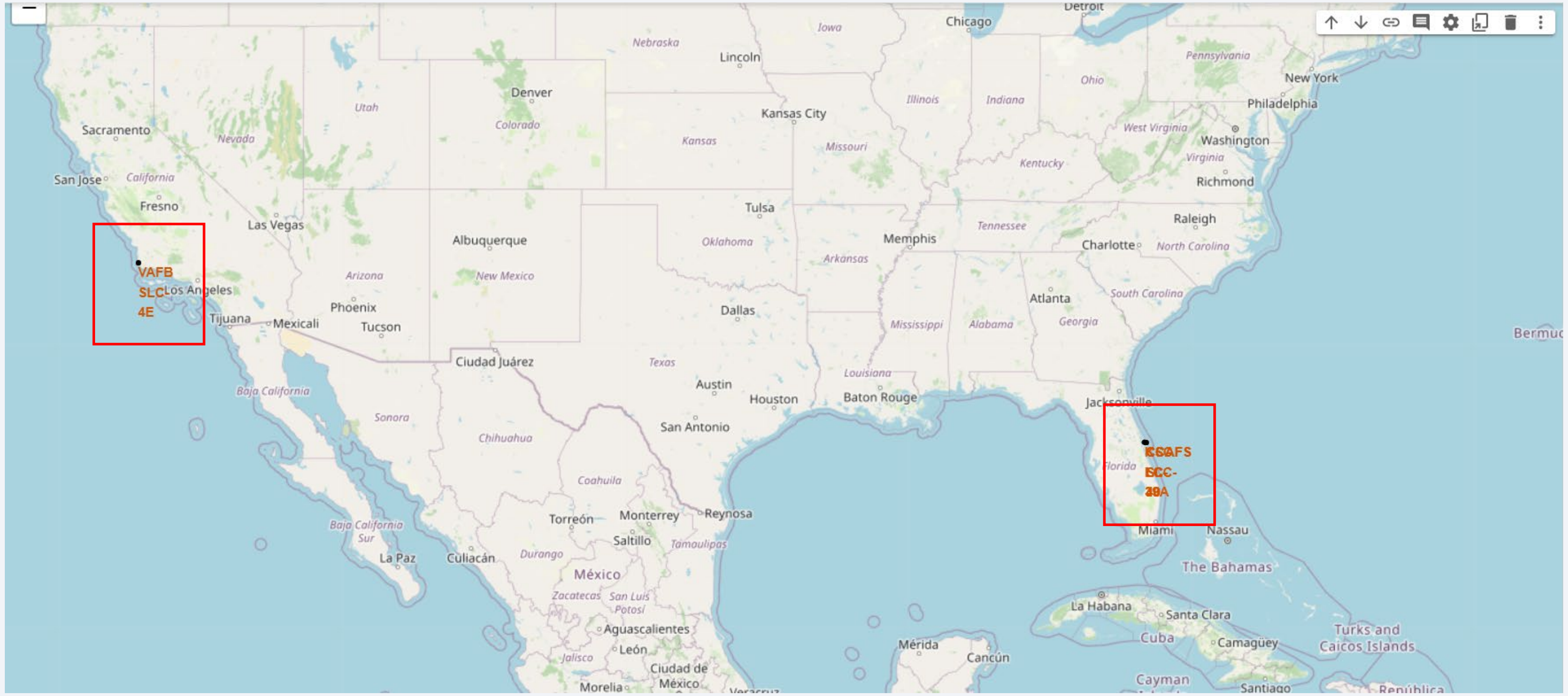
landing__outcome	COUNT OF LANDING OUTCOME
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

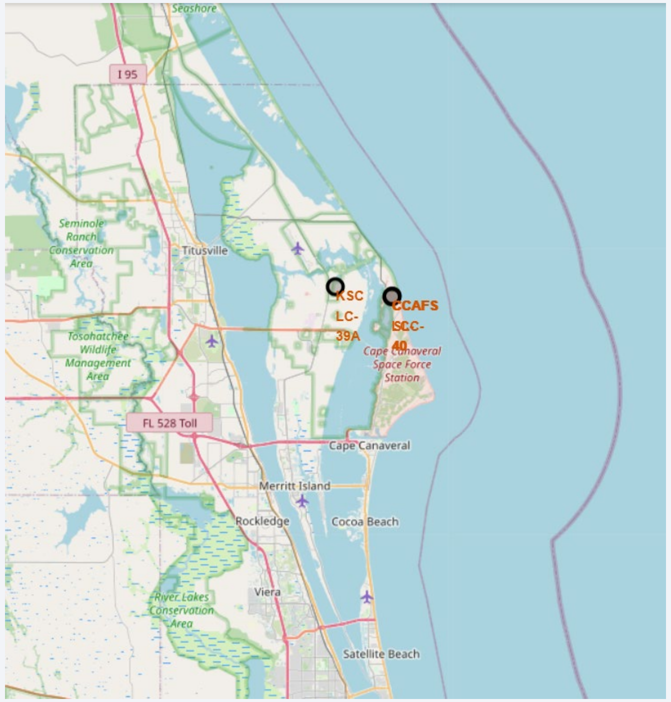
- Most landing attempts occur by drone ship (11 out of 31)
- Successful and failed attempts are just about equal (11 and 10, respectively)

Section 3

Launch Sites Proximities Analysis

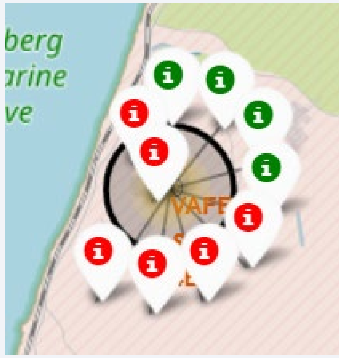
Launch Site Locations in Continental U.S.



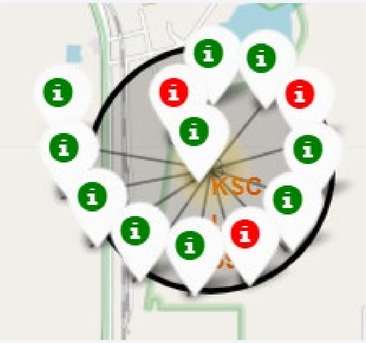


- All but one launch site in Florida
- VAFB in California
- All locations on coastline in southern part of U.S.

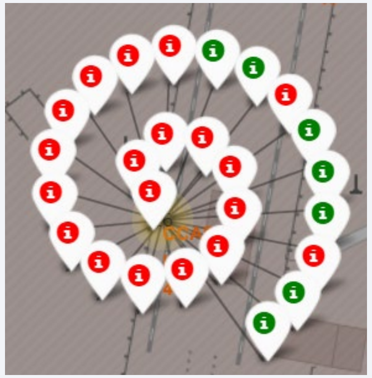
Success/Fail Launch Outcomes for Launch Sites



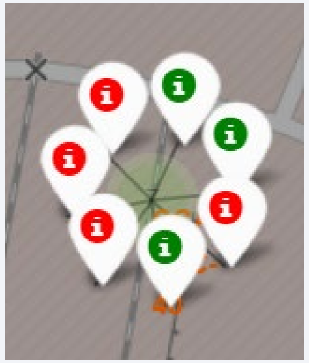
VAFB SLC-4E



KSC LC-39A



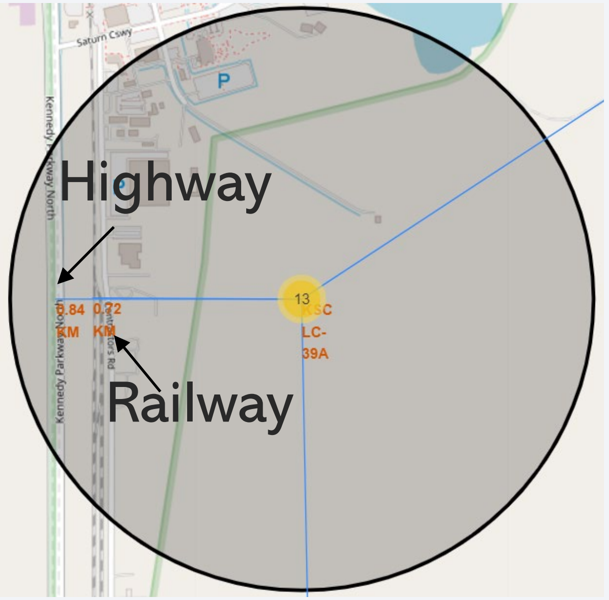
CCAFS LC-40

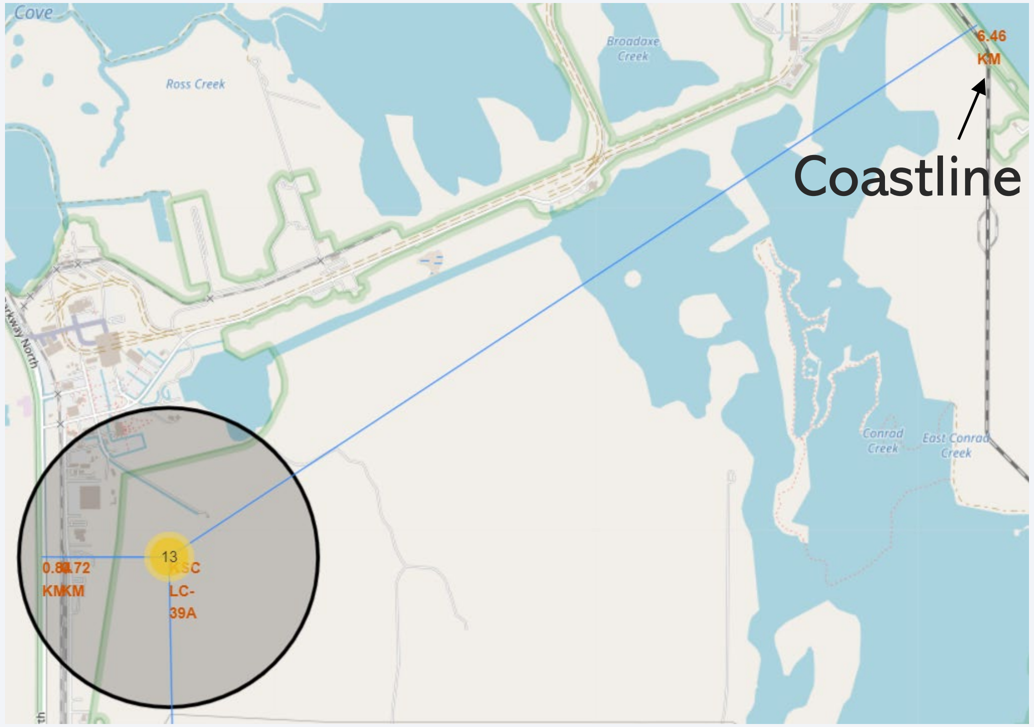


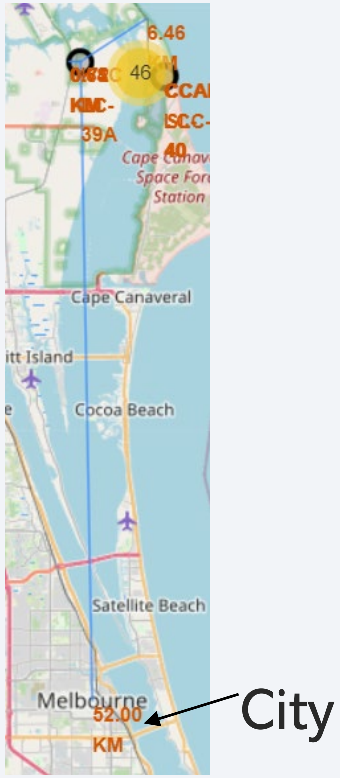
CCAFS SLC-40

- CCAFS yielded most flights (LC-40 and SLC-40)
- Despite being on the opposite coast as Florida, VAFB did not yield a significantly higher success rate (40%)
- KSC, despite being close to CCAFS, yielded much higher success rate

Landmark Distances from KSC SLC-39A



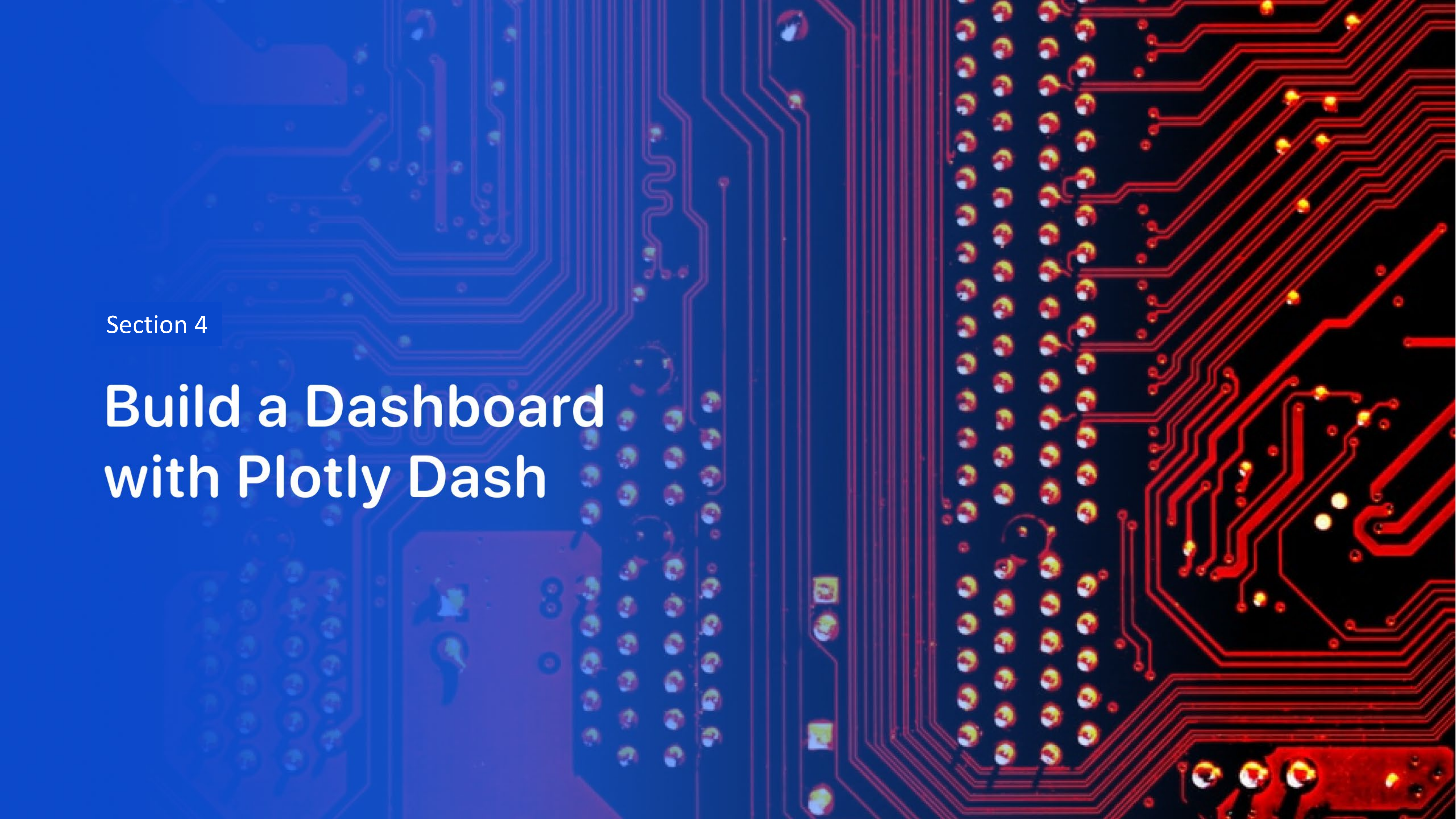




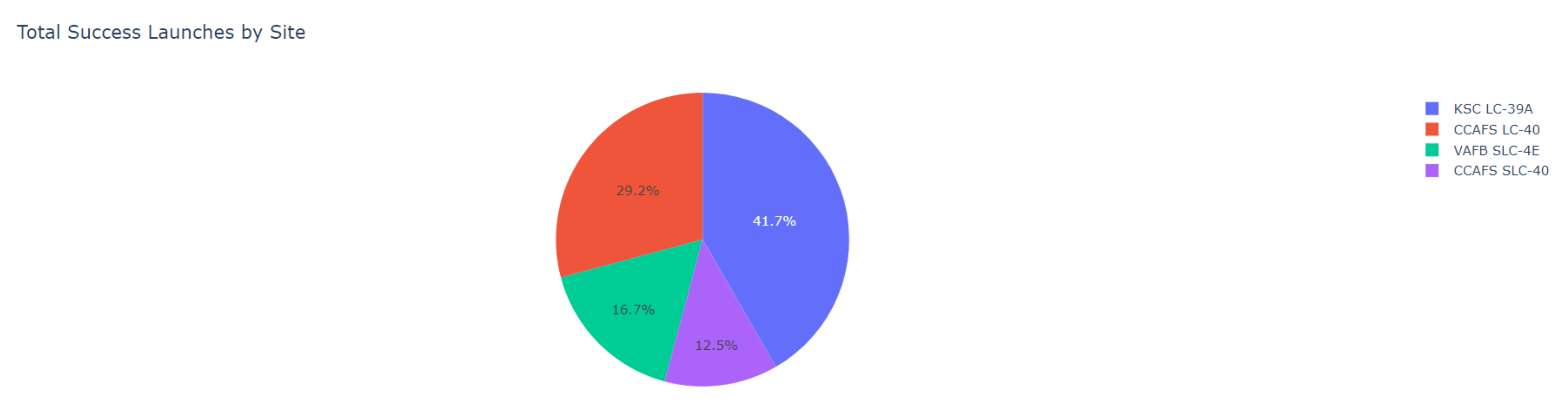
- Distances from railways and highways are short (<1 km.)
- Coastline further away from KSC (6.5 km.) than from other sites (<1.5 km., see appendix)
- Nearest city over 50 km away from launch site

Section 4

Build a Dashboard with Plotly Dash

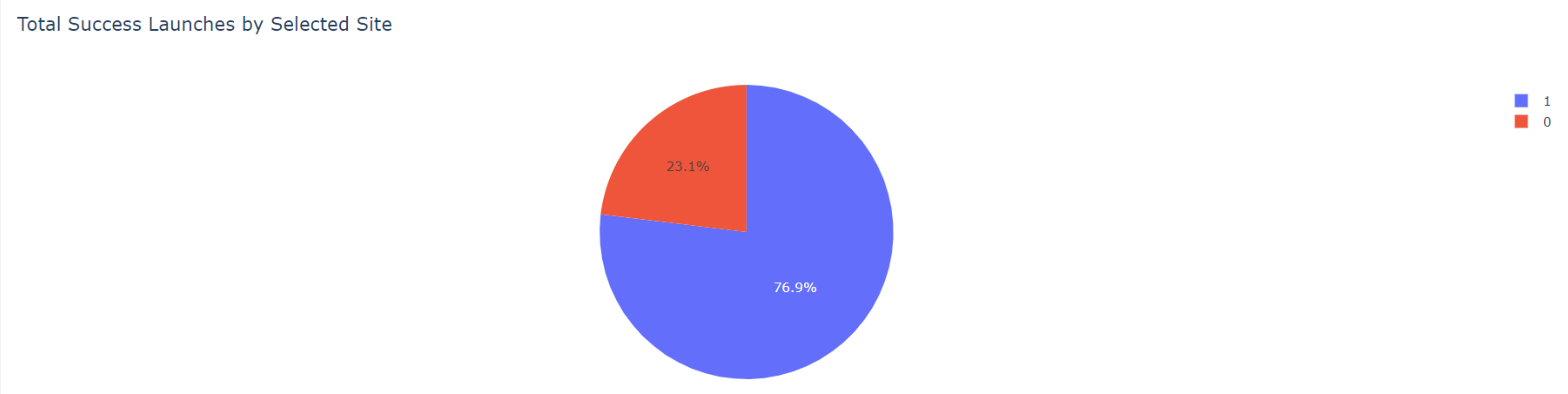


Distribution of Successful Launches by Site



- KSC launch site yielded most successful launches (over 40%)
- VAFB site yielded least successful launches
- CCAFS SLC-40 is same as LC-40, which accounted for 41.7% of successful launches

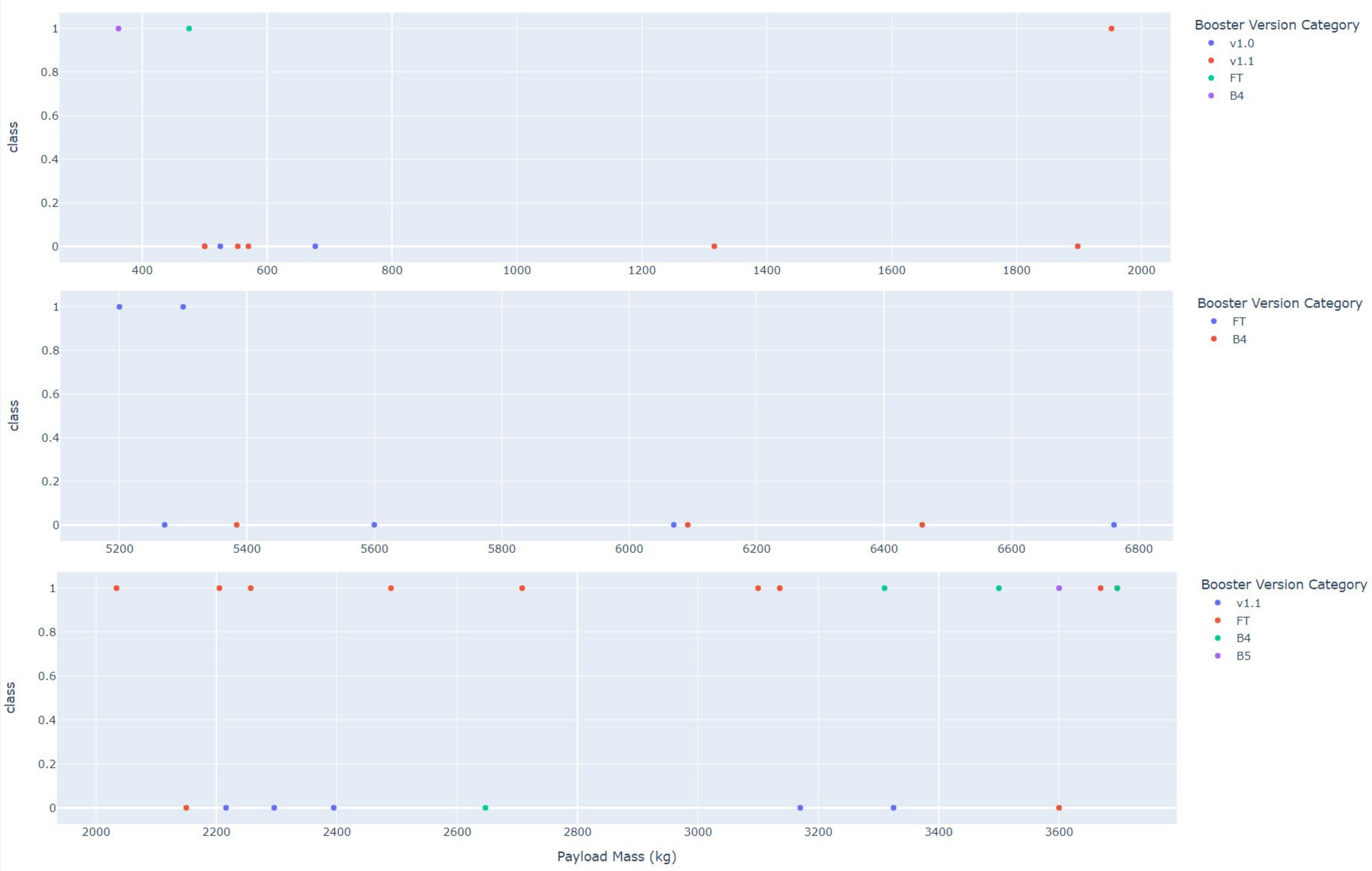
Success/Failure Rate at KSC Launch Site



- 76.9% of launches at KSC were successful, highest among all sites
- Correlates with highest number of successful launches compared to all sites (previous slide)
- All other sites yield a successful rate below 50%

Launch Outcome by Payload Mass Ranges and Booster Used

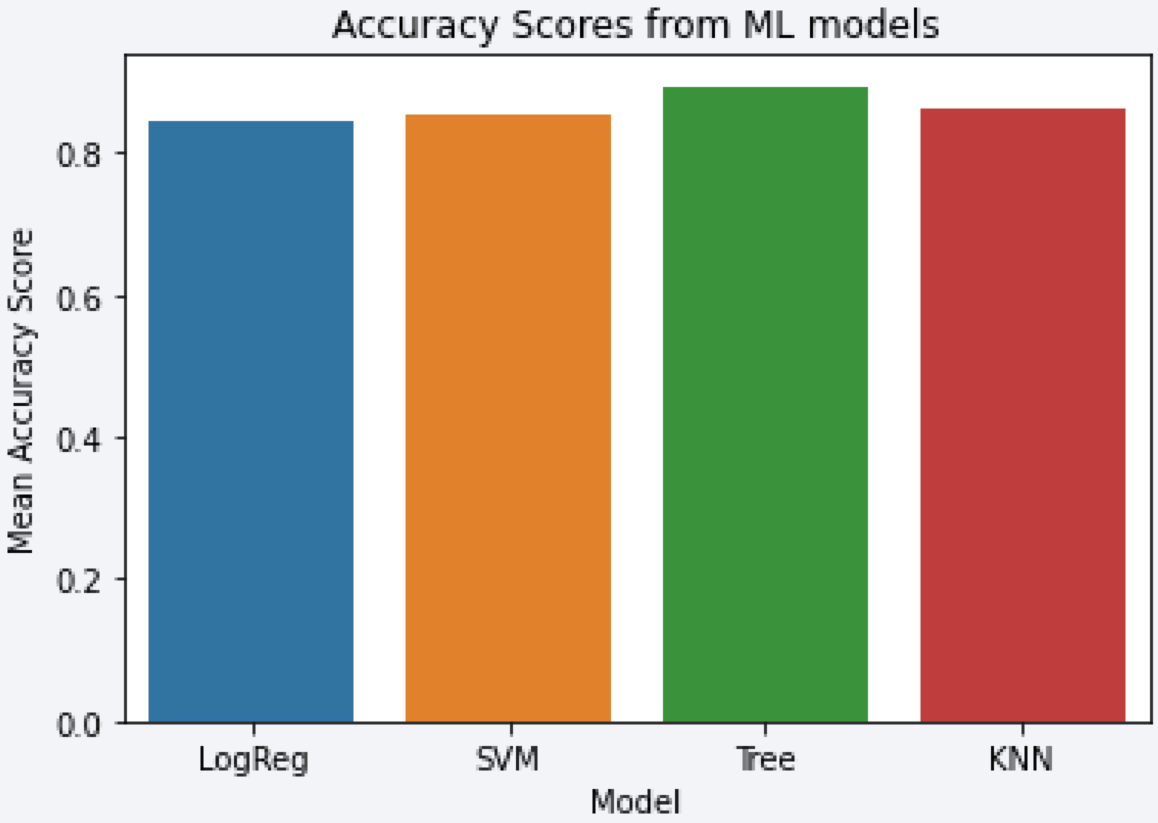
- Most failed launches carry <2000 kg. or 5000-7000 kg.
- Most successful launches in 2000-4000 kg.
- FT is most successful booster, prominently in 2000-5000 kg. range
- V1.1 is least successful booster



Section 5

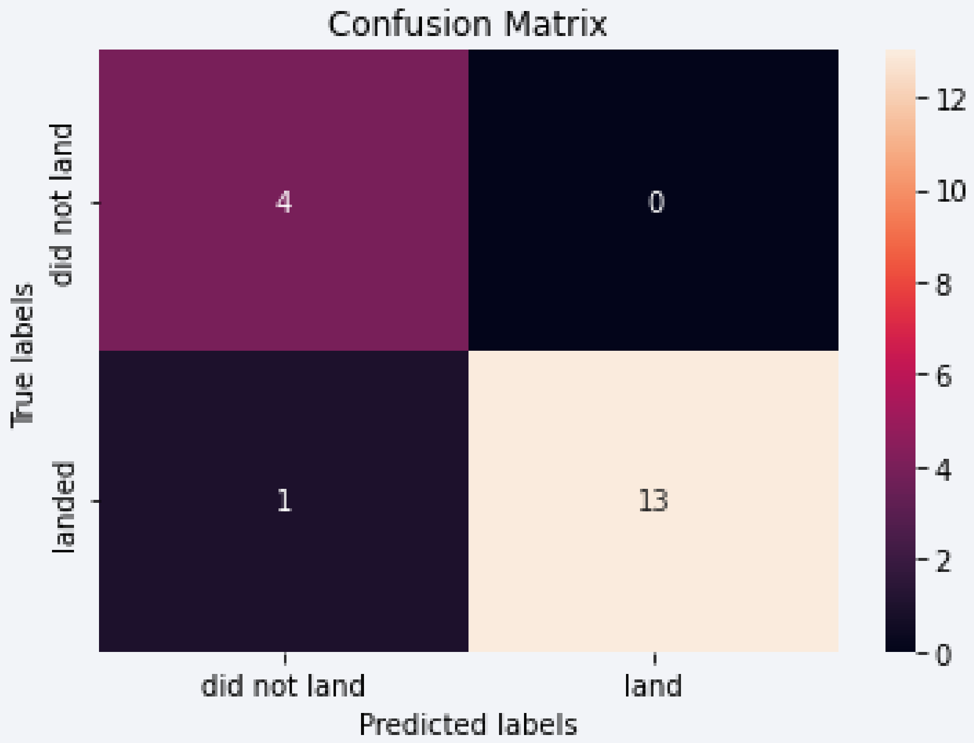
Predictive Analysis (Classification)

Classification Accuracy



Decision Tree yielded the highest average model accuracy across 5 iterations of training and testing.

Confusion Matrix of Decision Tree Classifier



- No false positives, only one false negative in this iteration of model
- Better to err on the side of mislabeling as failure than mislabeling as success
- Test score of 0.944

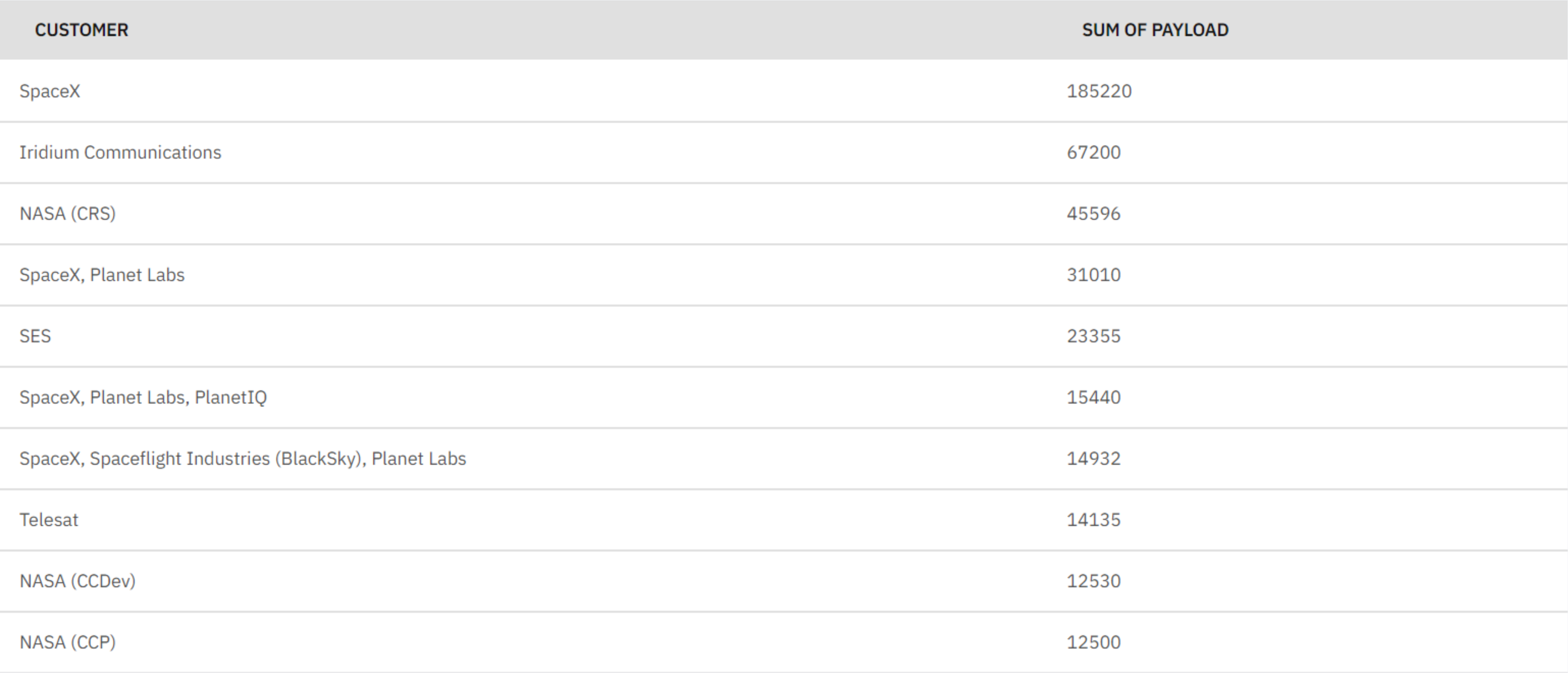
Conclusions

- Heavier payload mass generally yielded greater success
  - Different orbit types have specific ranges for success (Low payload mass for LEO, SSO, MEO; high payload mass for VLEO)
- F9 FT boosters best for mid-range payload masses, F9 B5 best for high-range payload masses
  - Avoid earlier booster versions (F9 v1.0, F9 v1.1)
- Launch sites should be close to coastlines and far away from cities (>50 km.)
  - Railways and highways may remain close to launch sites (<1 km.)
- Decision Tree model yields best classification, despite slight variations between iterations

Section 6

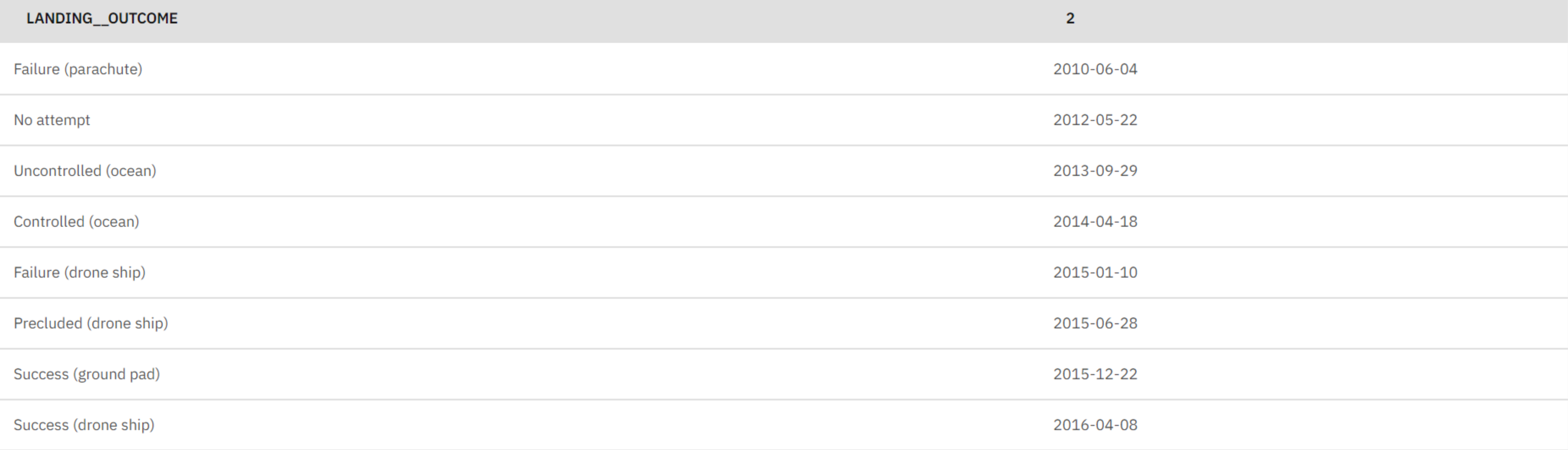
Appendix

Top Customers of Falcon 9 Launches by Payload Mass



CUSTOMER	SUM OF PAYLOAD
SpaceX	185220
Iridium Communications	67200
NASA (CRS)	45596
SpaceX, Planet Labs	31010
SES	23355
SpaceX, Planet Labs, PlanetIQ	15440
SpaceX, Spaceflight Industries (BlackSky), Planet Labs	14932
Telesat	14135
NASA (CCDev)	12530
NASA (CCP)	12500

Earliest Date for Each Landing Outcome

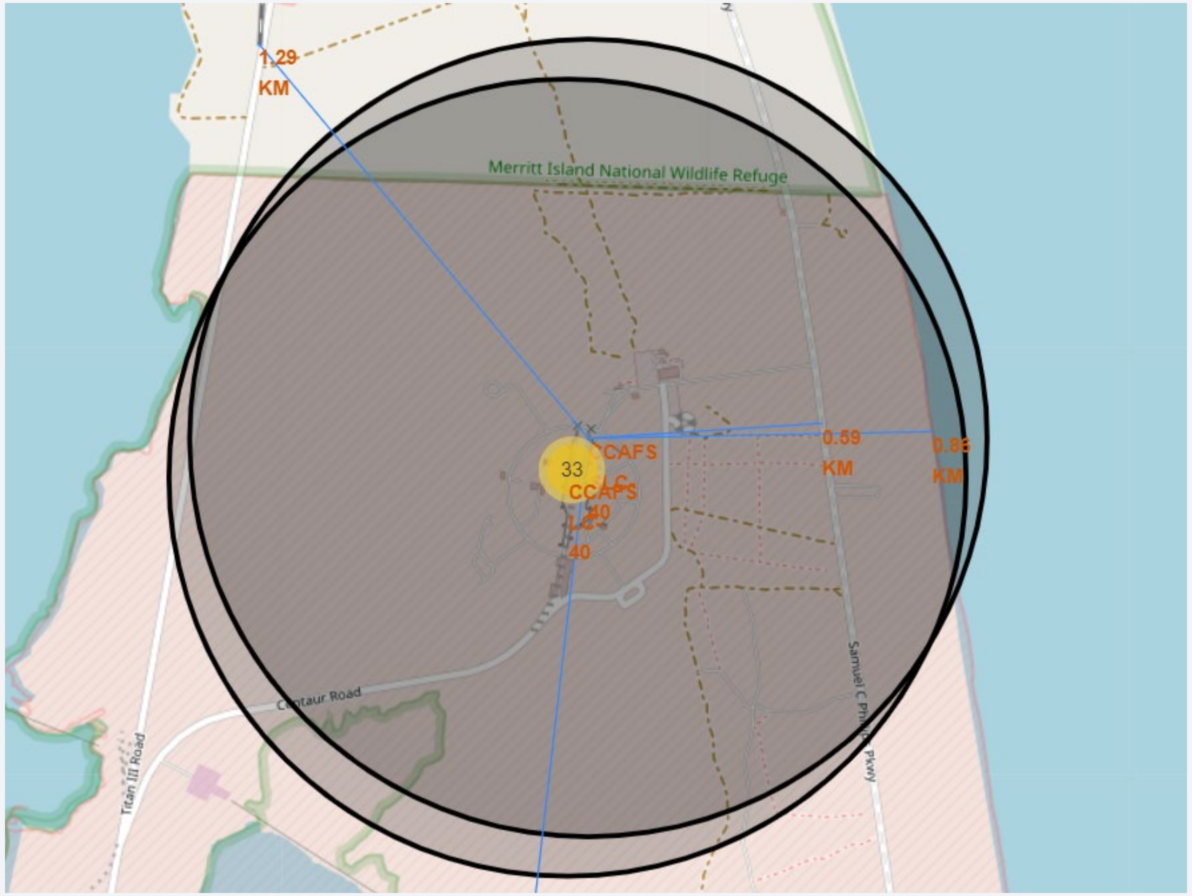


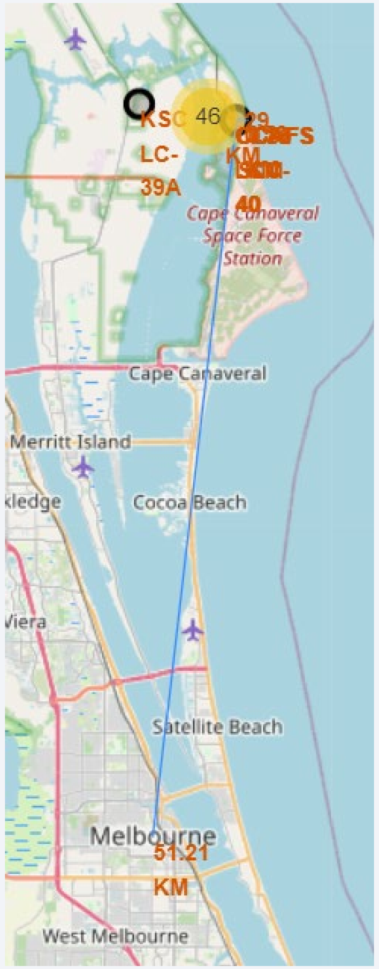
LANDING__OUTCOME	2
Failure (parachute)	2010-06-04
No attempt	2012-05-22
Uncontrolled (ocean)	2013-09-29
Controlled (ocean)	2014-04-18
Failure (drone ship)	2015-01-10
Precluded (drone ship)	2015-06-28
Success (ground pad)	2015-12-22
Success (drone ship)	2016-04-08

Average Payload Mass by Booster Version

Booster Version	Average Payload Mass (kg)
F9 v1.0	340
F9 v1.1	2534
F9 B4	4970
F9 B5	9132
F9 FT	4568

Landmark Distances from CCAFS SLC-40

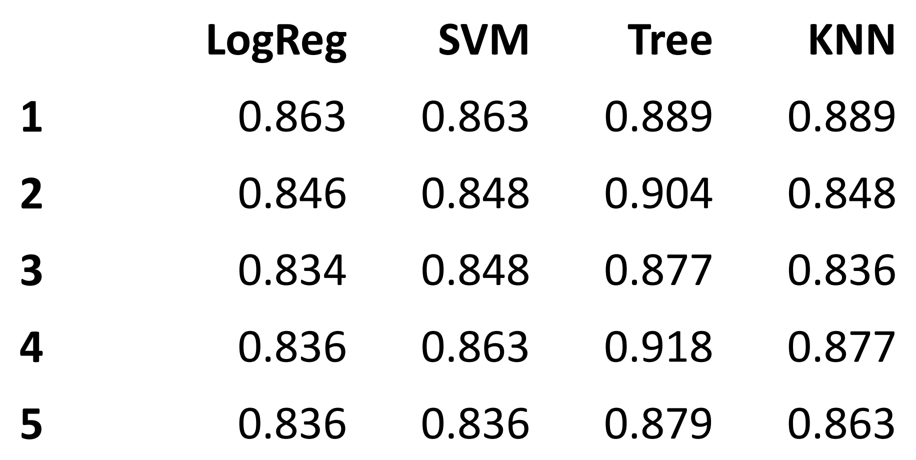




Scores from All Classification Models

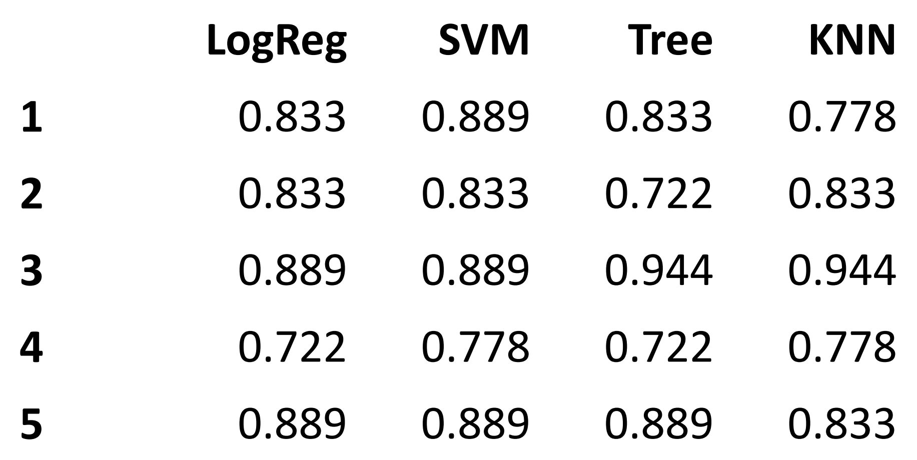
Accuracy (from Training Set)

Iteration



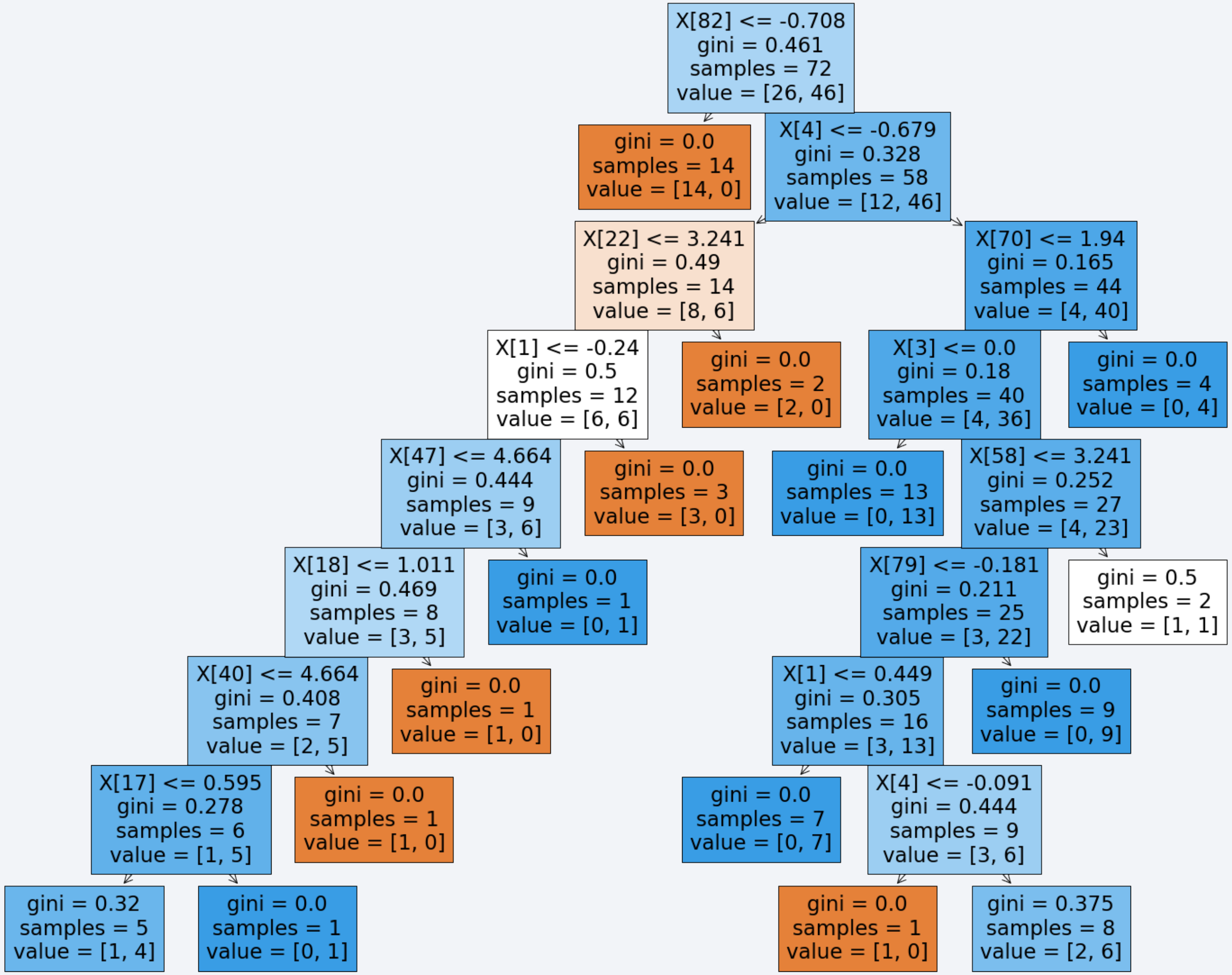
	LogReg	SVM	Tree	KNN
1	0.863	0.863	0.889	0.889
2	0.846	0.848	0.904	0.848
3	0.834	0.848	0.877	0.836
4	0.836	0.863	0.918	0.877
5	0.836	0.836	0.879	0.863

Classification (from Test Set)



	LogReg	SVM	Tree	KNN
1	0.833	0.889	0.833	0.778
2	0.833	0.833	0.722	0.833
3	0.889	0.889	0.944	0.944
4	0.722	0.778	0.722	0.778
5	0.889	0.889	0.889	0.833

Decision Tree Model



Thank you!