Winning Space Race with Data Science

Jacob Contreras
14 October 2021

Outline

Executive Summary

- Launch data was collected through two means:
  - Making requests to the SpaceX REST API
  - Webscrpaing a Wikipedia article of launch data using the BeautifulSoup library
- Data was processed into values that could be used for visualization and modeling
- Since our prediction result is binary (success or failure), different classification models were trained and tested to find which was the best. Among K Nearest Neighbors, Logistic Regression, Support Vector Machines, and Decision Trees, it was determined that all except for decision trees produced a similar accuracy that was also the max accuracy of the four models.
- Analysis of the data gave insights into the importance of factors like the flight number (higher flight numbers correlated with higher rates of success) and orbit type (with SSO, HEO, GEO, ES-L1 having the highest success rates).

Introduction

- Ingenuity from SpaceX's Falcon 9 rocket has allowed the company to conduct launches at a lower cost than its competitors. However, this lowered cost is entirely dependent upon their ability to successfully land the first stage of the rocket.
- Our goal is to predict whether an upcoming Falcon 9 launch will successfully land its Stage 1, or if SpaceX will lose money from a failed landing.
- We will find out what factors are common to successful landings and build a model to predict outcomes for future launches.

Section 1

Methodology

Methodology

- Executive Summary
- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

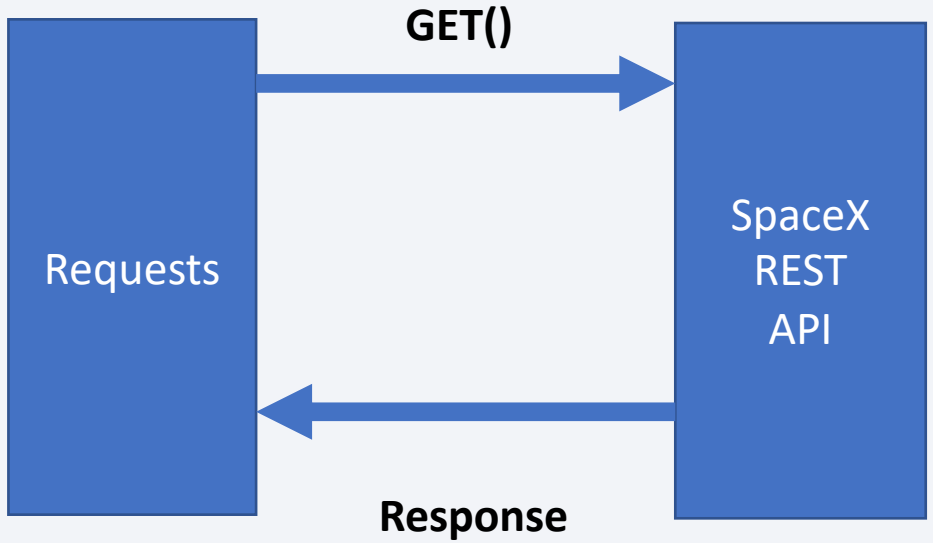
- Data was collected by two means: a REST API for SpaceX data and by web scraping a Wikipedia article about past SpaceX launches.
- For the REST API, the Requests Python library was used to make data requests using a get request. We can then check the API's response for the JSON file containing the data requested.
- For webscraping, the Requests library was used to grab the html of the webpage, and the BeautifulSoup library was used to parse the html.

Data Collection – SpaceX API

To access information from the SpaceX REST API, the Requests library was used.

Several GET calls were made to fetch JSON files, which were then parsed by defined python functions in the Jupyter Notebook.

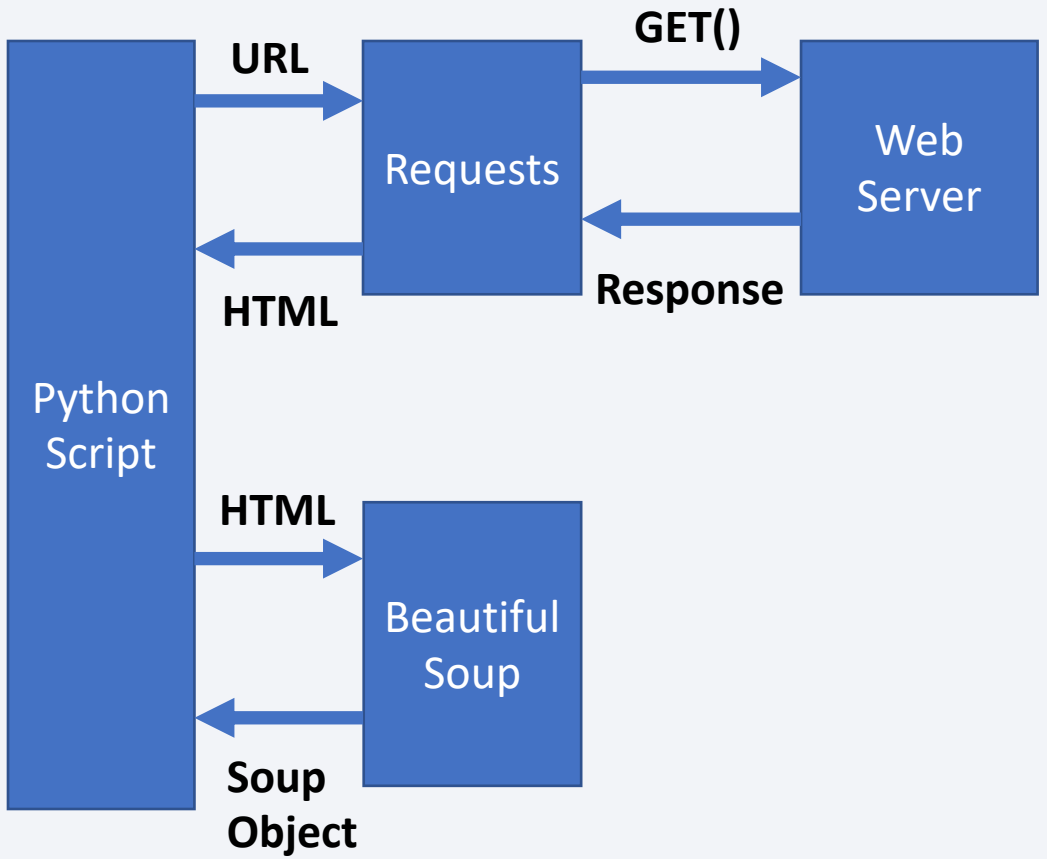
https://github.com/JackCon21/Coursera_Capstone-SpaceX/blob/master/SpaceX%20Data%20Collection.ipynb



Data Collection - Scraping

- Data was accessed from a Wikipedia article on SpaceX Falcon 9 launches.
- Using Requests, we GET the html of the webpage. The html is then parsed using BeautifulSoup. Beautiful soup can then be used to find and extract tabular data.

https://github.com/JackCon21/Coursera_Capstone-SpaceX/blob/master/SpaceX%20-%20Data%20Collection%20(Web%20Scraping).ipynb



Data Wrangling

- Data was processed using the Pandas and Numpy libraries
- The element we wanted to predict was the outcome of the landing. The initial column for outcomes contained object types and we wanted a classification of success or failure. We collected a list of outcomes that signified failure and used this list to convert all failures to 0 signifying a failed outcome and a 1 for a success.

https://github.com/JackCon21/Coursera_Capstone-SpaceX/blob/master/SpaceX%20-%20Data%20Wrangling.ipynb

EDA with Data Visualization

- Data was visualized using a collection of bar graphs, scatter plots, and line plots.
- Since the outcome we are predicting is categorical, bar graphs and scatter plots were useful in plotting different attributes against the categorical outcome of success or failure. A third dimensional attribute of scatter plots, hue, was helpful in distinguishing the success or failure.
- A line plot was useful in showing the trends of success rate over the course of time with increasing launch number.

https://github.com/JackCon21/Coursera_Capstone-SpaceX/blob/master/SpaceX%20-%20EDA%20(Data%20Visualization).ipynb

EDA with SQL

- SQL queries were used to:
  - Find Unique Launch Sites
  - Find the unique outcomes of launched
  - Explores average and total payload masses
  - Determine which boosters carried the maximum payload

https://github.com/JackCon21/Coursera_Capstone-SpaceX/blob/master/SpaceX%20-%20Exploratory%20Data%20Analysis%20(SQL).ipynb

Build an Interactive Map with Folium

- A folium map was created with:
  - A circle and marker for each launch site
  - Marker Clusters containing markers for all launches, successful and unsuccessful, at each site
  - Lines from the VAFB to the nearest coast, railroad, highway, and city with the Euclidean distance between each location and the launch site

https://github.com/JackCon21/Coursera_Capstone-SpaceX/blob/master/SpaceX%20-%20Visual%20Analytics%20(Folium).ipynb

Build a Dashboard with Plotly Dash

- The SpaceX Launch Records Dashboard contains
  - A dropdown to filter which data is displayed. The dropdown includes each launch site as well as an ALL option to show overall launch outcomes
  - A range slider to set the filter data to be within a payload mass range
  - A pie chart, displaying the success rate of landings.
  - A scatter plot, displaying landing outcomes vs the payload range established by the range slider.

https://github.com/JackCon21/Coursera_Capstone-SpaceX/blob/master/spacex_dash_app.py

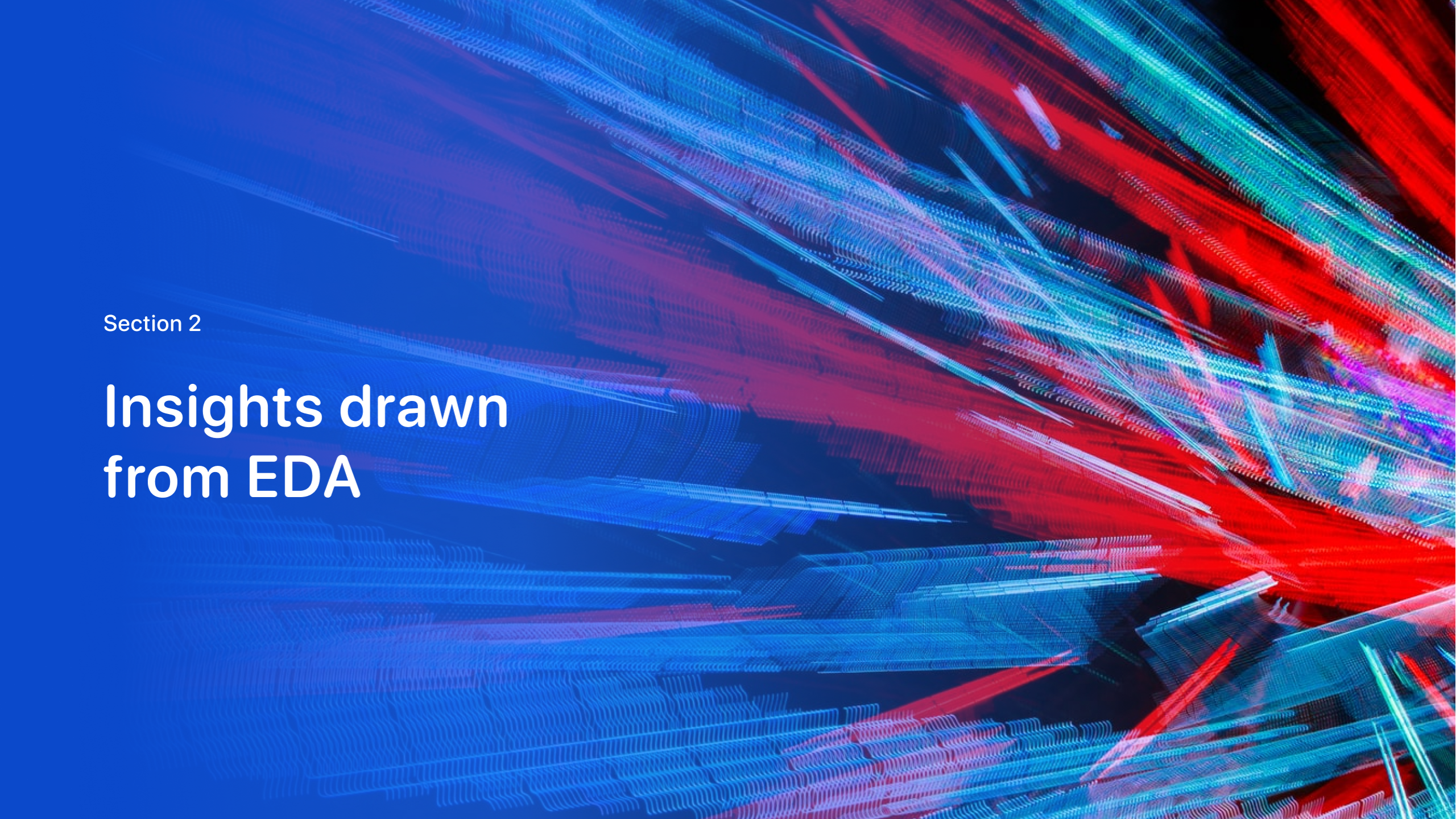
Predictive Analysis (Classification)

- The goal of the predictive model was to decide if a Falcon 9 stage 1 landing was successful or a failure. Since this outcome is categorical, a classification model would be the best fit.
- Models for K Nearest Neighbors, Support Vector Machine, Logistic Regression, and Decision Trees were created to see which was the best predictor.
- The data was split into a testing set and a training set and normalized.
- Each model was fit with the training data and then tested with the test data to generate an accuracy score.

https://github.com/JackCon21/Coursera_Capstone-SpaceX/blob/master/SpaceX%20-%20Predictive%20Analysis.ipynb

Results

- From the Exploratory Data Analysis, it was determined that the flight number was an important indicator in the success of a flight (as the flight number increased so did the success rate). It was also determined that SSO, HEO, GEO, and ES-L1 orbital flights had the highest success rate.
- From the Interactive analysis, successful landings occurred most often in when carrying a payload between 2000 and 6000 kg. It was also determined that launches from Kennedy Space Center had the highest success rate among all four sites where Falcon 9 launches have taken place.
- Predictive analysis produced generally high accuracy models. The decision tree model was the worst performing with the out of sample data with ~77% accuracy. The other three models were equally as accurate as each other with ~83% accuracy.

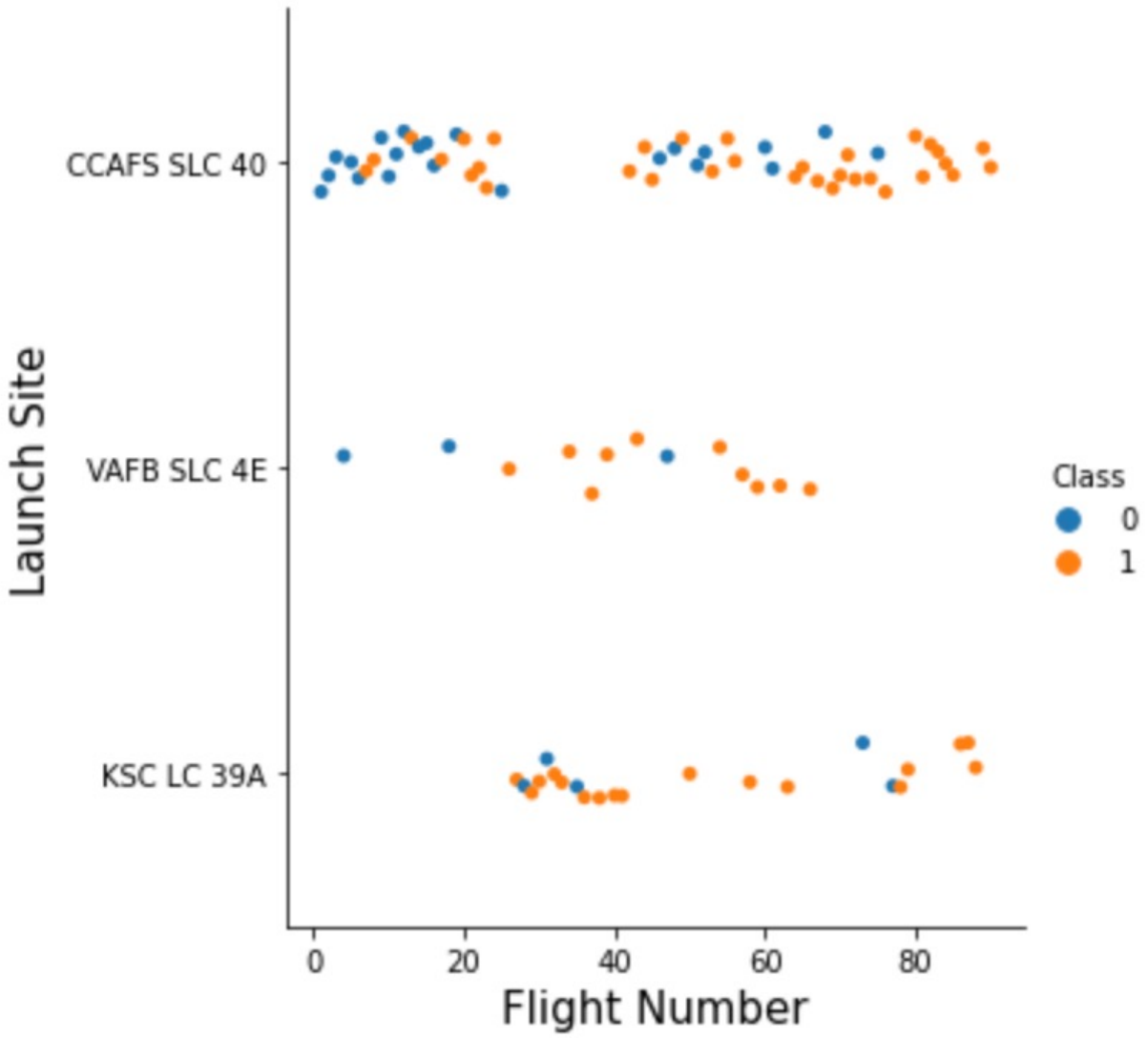


Section 2

Insights drawn from EDA

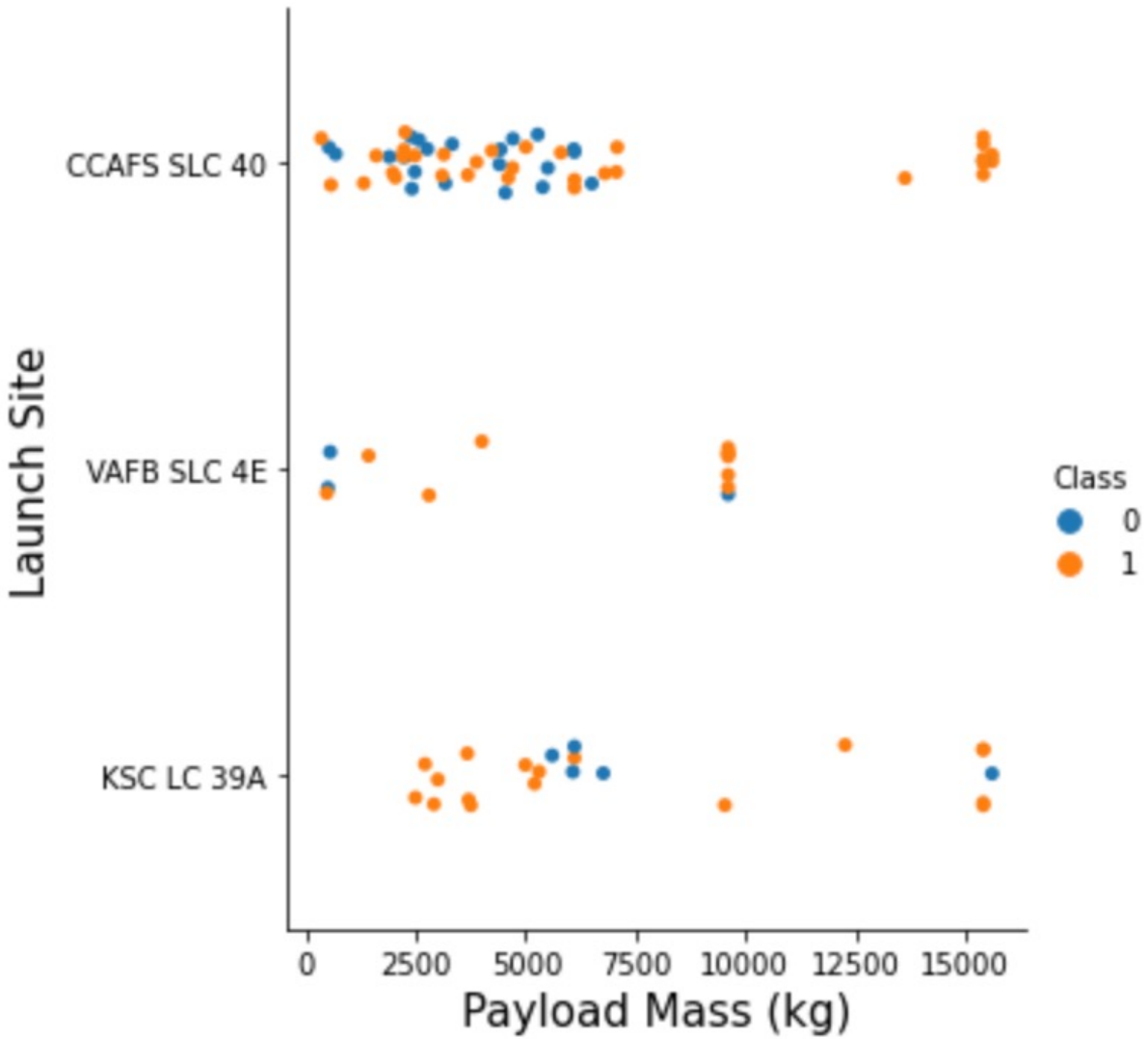
Flight Number vs. Launch Site

- VAFB SCL 4E shows an improvement in successes as the flight number increases.
- The density of failures at CCAFS SLC 40 also diminishes as the flight number increases.



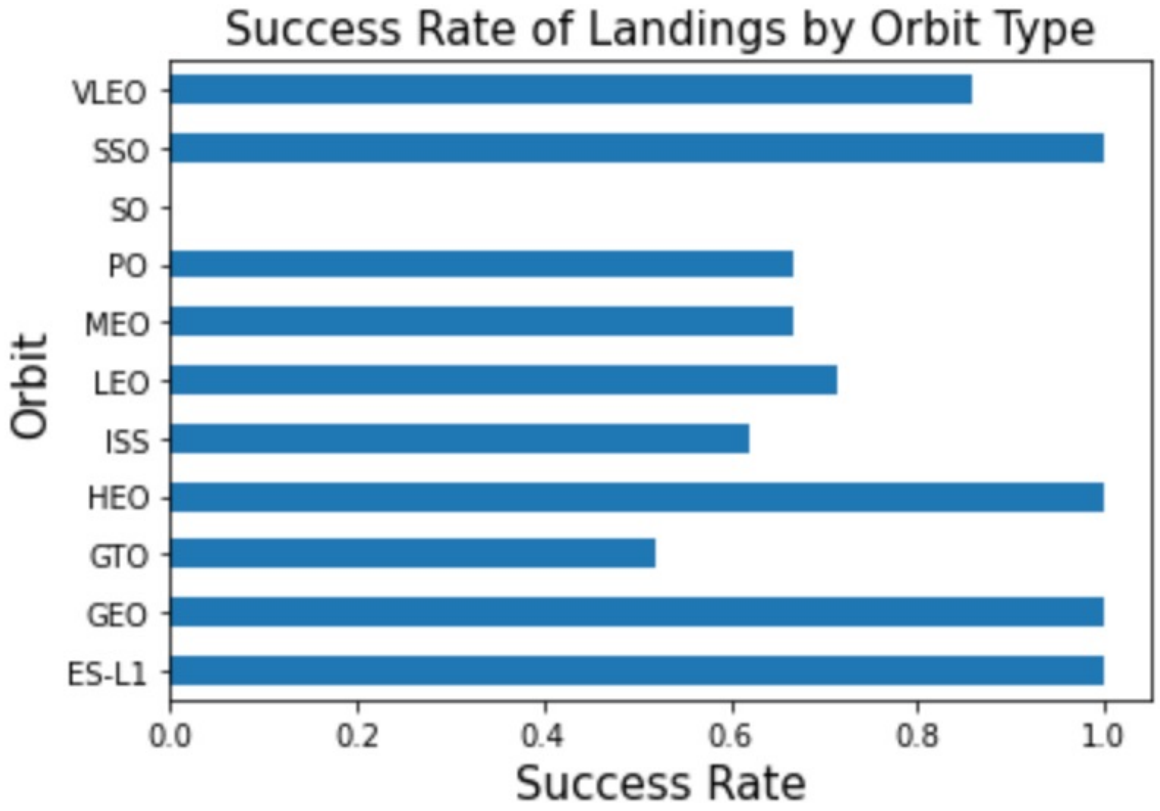
Payload vs. Launch Site

- At the CCAFS SLC 40 site, heavier payloads had high success.
- Low weight payloads at the same site are inconclusive about whether the mass matters.
- KSC LC 39A has a high success between 2500 and 5000 kg.



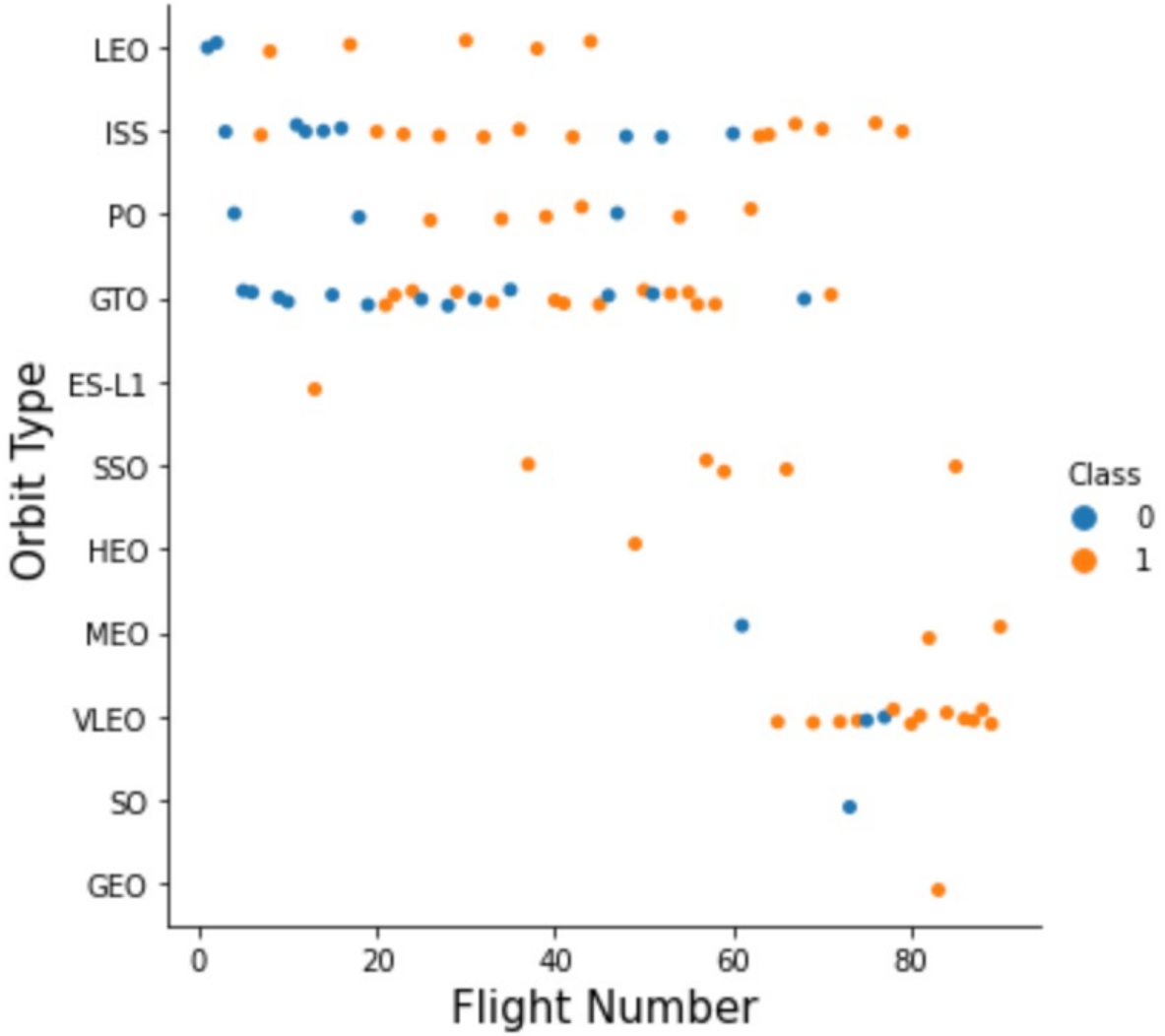
Success Rate vs. Orbit Type

- SSO, HEO, GEO, and ES-L1 Orbits had the highest success rate.
- SO may be an outlier to the data as SSO and SO are synonymous



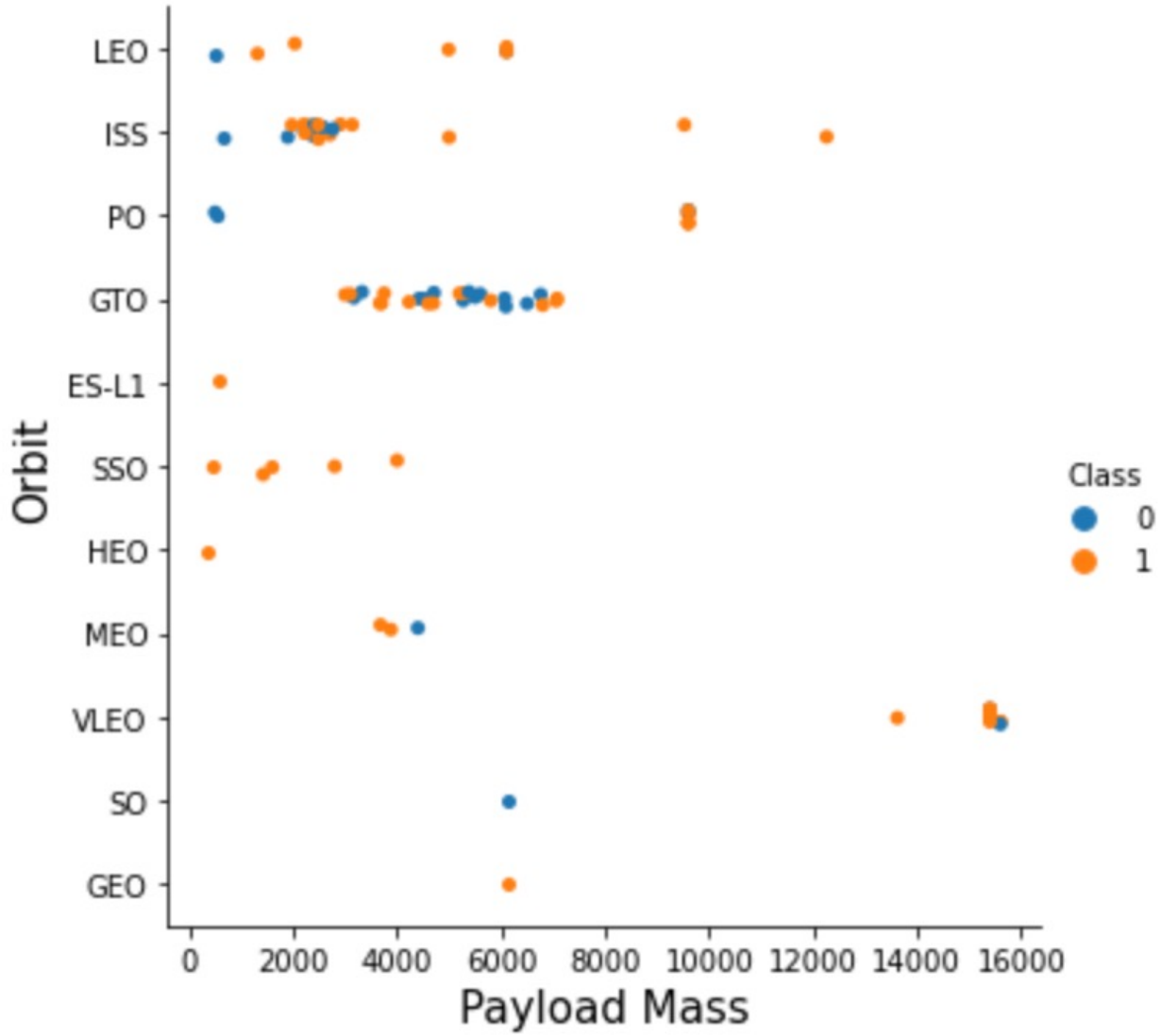
Flight Number vs. Orbit Type

- Higher flight numbers tend to have a higher rate of success. This trend is present in the LEO orbit type.



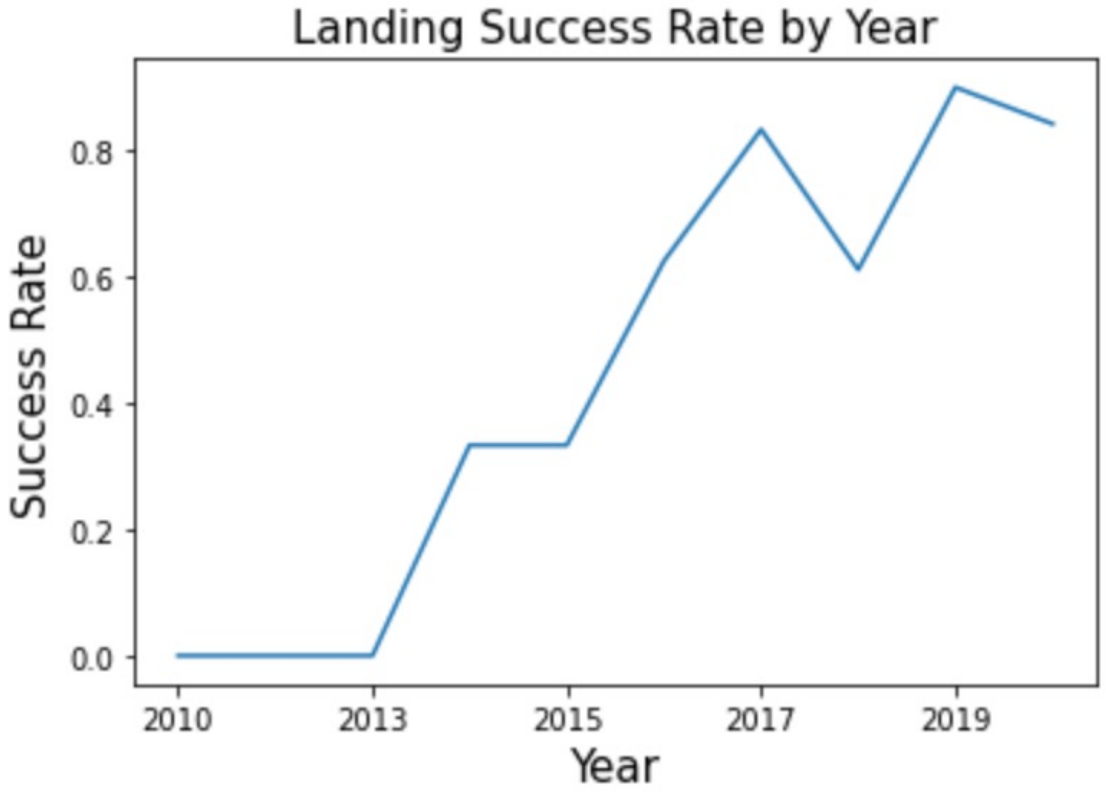
Payload vs. Orbit Type

- Orbit types LEO, ISS, and PO have higher success rates with heavier payloads.
- Conclusions cannot be made about GTO as the data points appear to be evenly split between success and failure.



Launch Success Yearly Trend

- Success rate appears to be correlated with time.
- As time progresses the average yearly success rate tends to get higher.



All Launch Site Names

- There are four launch sites where F9 rockets were launched.
- 2 at Cape Canaveral Air Force Station
- Kennedy Space Center
- Vandenburg Air Force Base

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

Launch Site Names Begin with 'CCA'

- The launch sites at Cape Canaveral Air Force Station are the only sites that contain CCA as seen in the previous query.
- So, this query will collect all launches that took place at Cape Canaveral.
- The image to the right shows the first 5 launches at Cape Canaveral.

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

Total Payload Mass

total_mass_nasa
45596

Launches where the customer was NASA had a combined Payload mass of 45596 kg.

Average Payload Mass by F9 v1.1

average_mass
2534

The F9 v1.1 on average carried a payload mass of 2534 kg.

First Successful Ground Landing Date

first_success_ground
2015-12-22

SpaceX's first successful ground landing occurred December 12, 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

- The four boosters on the right all had successful landings on drone ships. They also carried payloads between 4000 and 6000 kg.

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

Total Number of Successful and Failure Mission Outcomes

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

- An overwhelming majority of missions were classified as being successful.
- Only one mission was deemed a failure outright.

Boosters Carried Maximum Payload

- 12 unique boosters have been used to carry the highest payload launched by SpaceX
- All 12 boosters are prefixed with F9 B5. So, they could all be similar models to each other.

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

- In 2015 only two landings to a drone ship were classified as failures.
- Both failures occurred at the same site Cape Canaveral Air Force Station
- They both appear to be a similar booster version having the prefix F9 v1.1.

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

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

- During the time period, all ground pad landings were successful
- The success rate of drone ship landings is 50%, with 5 successes and 5 failures
- A plurality of launches had no attempt at a landing

landing__outcome	total
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

Section 4

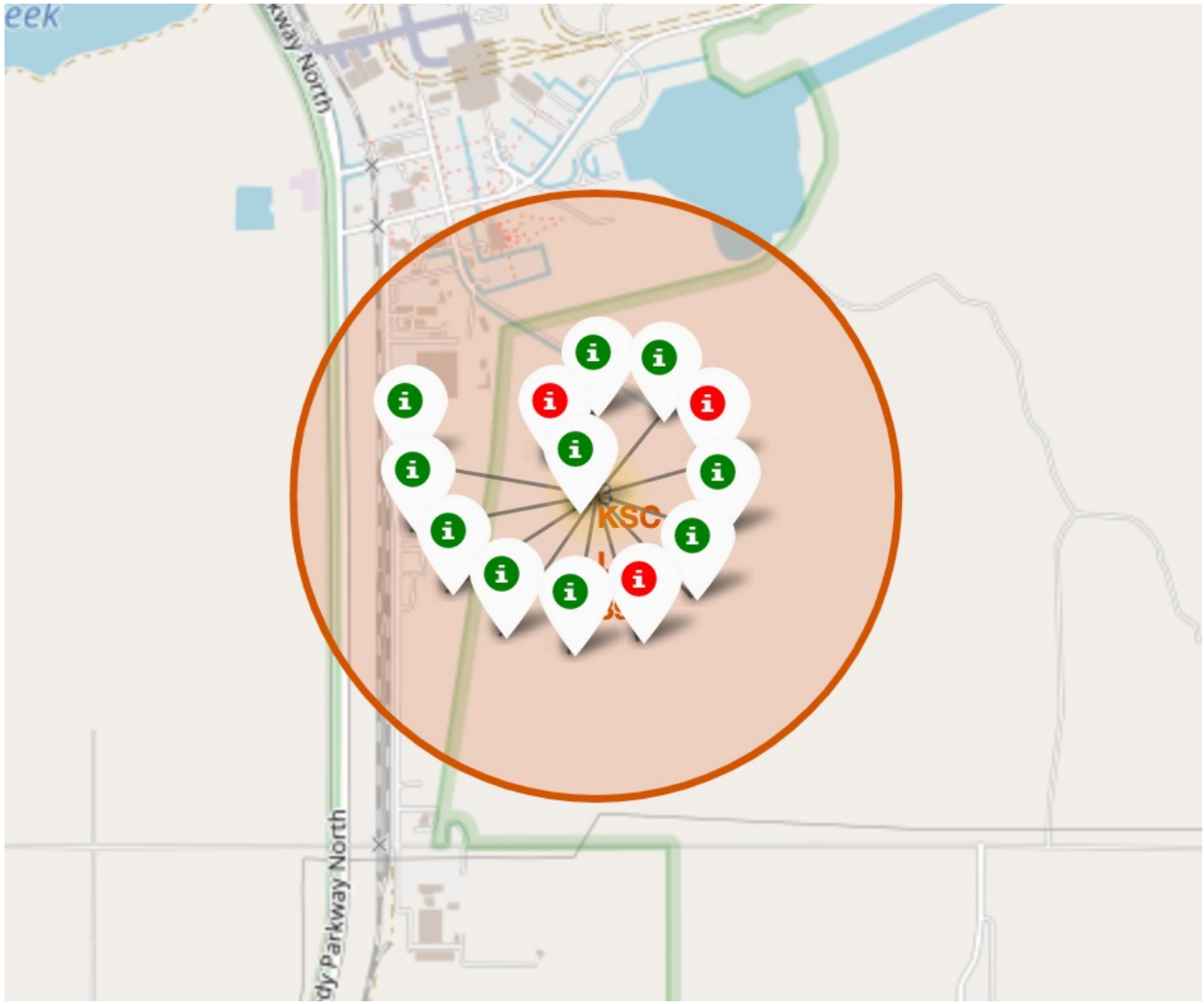
Launch Sites Proximities Analysis

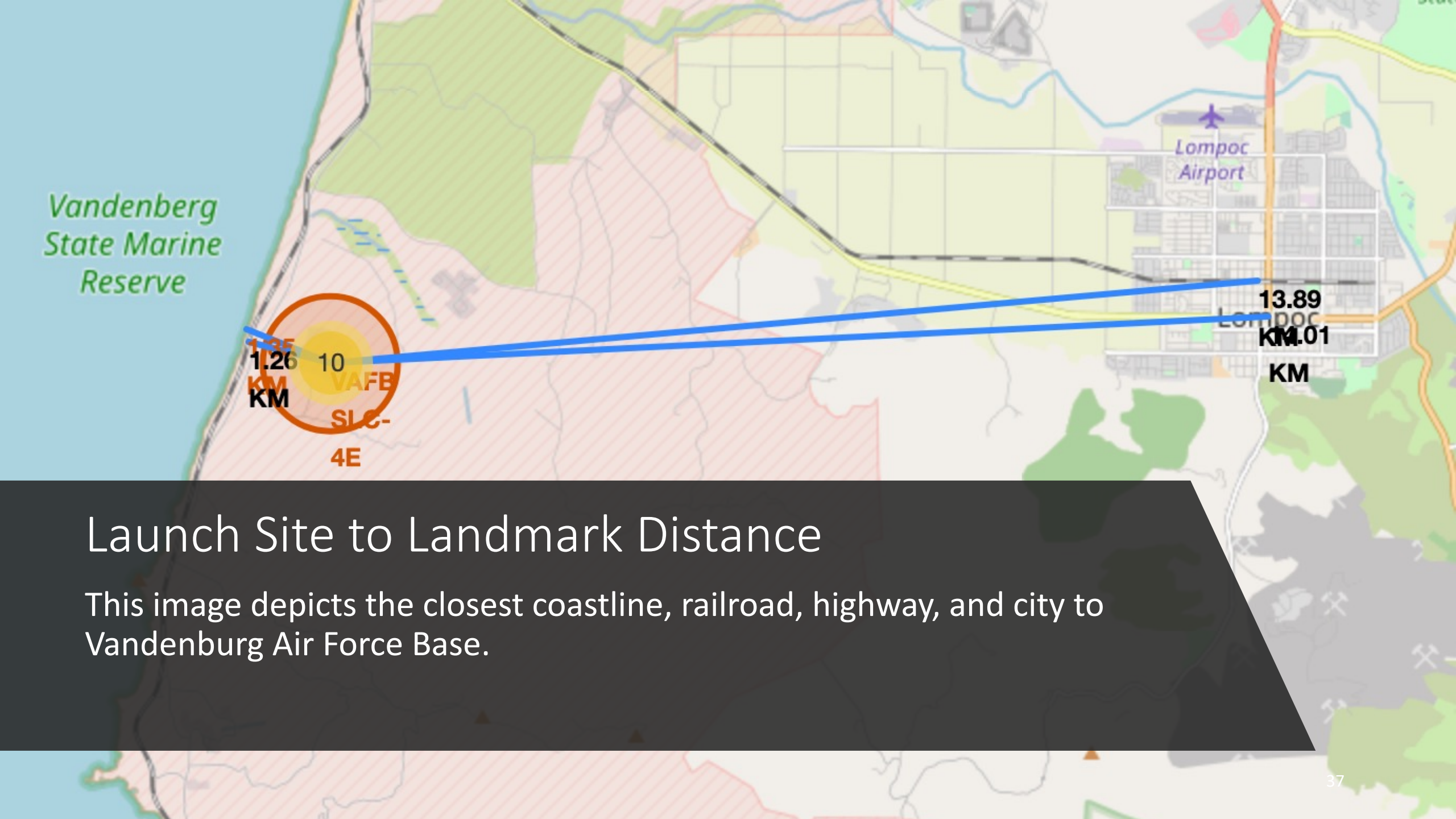
SpaceX Launch Sites

- Four launch sites are mapped above: VAFB SLC 4E, CCAFS SLC 40, CCAFS LC 40, and KSC LC 39A
- All sites are located on the coasts and are situated in the southern half of the United States.

Launch Outcomes

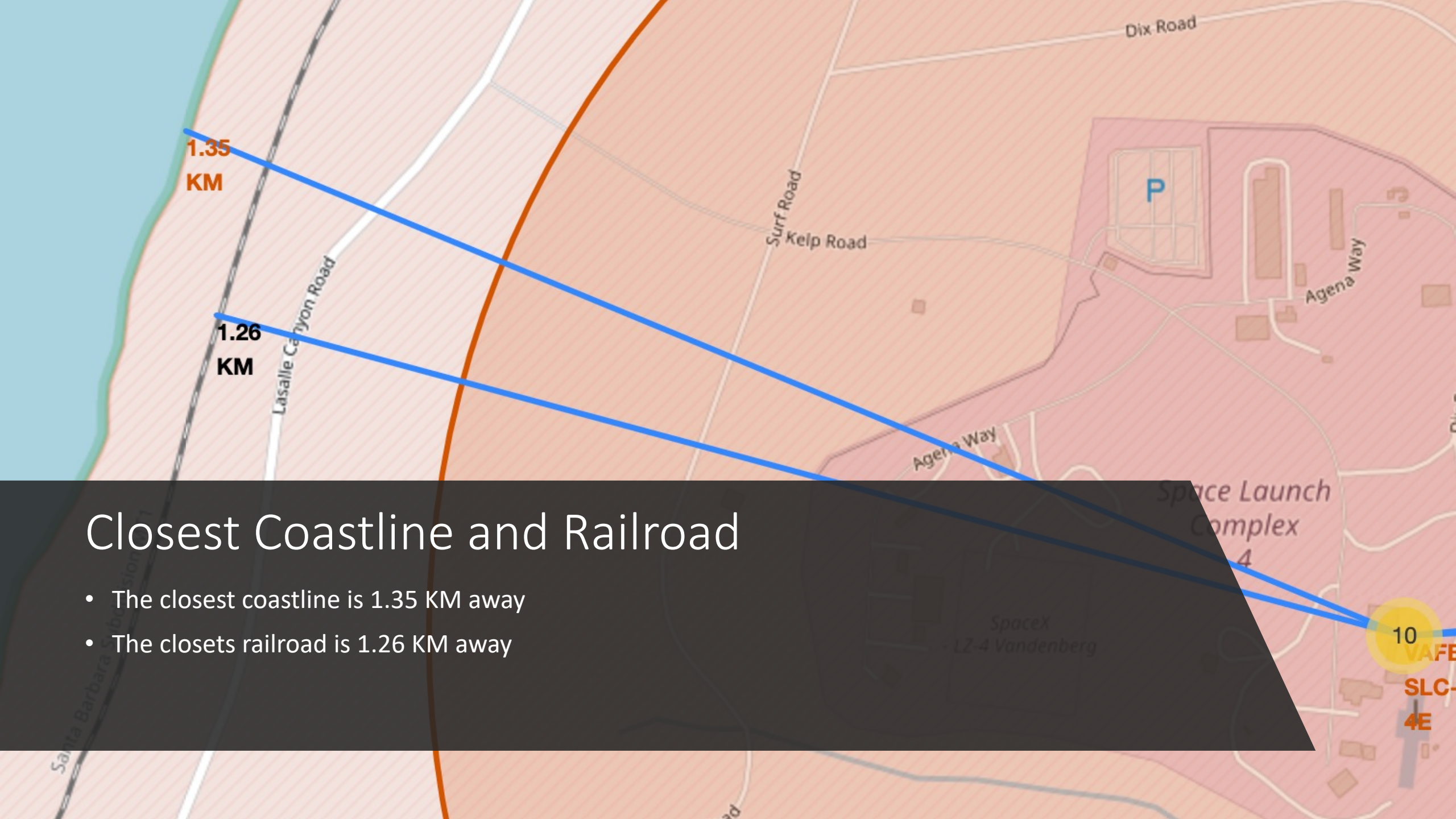
- Each launch site contains a marker cluster of launch outcomes at that site.
  - Red – Failure
  - Green – Success
- This image shows the launch outcomes for Kennedy Space Center.
- Far more successful landings occurred at KSC than failed landings.





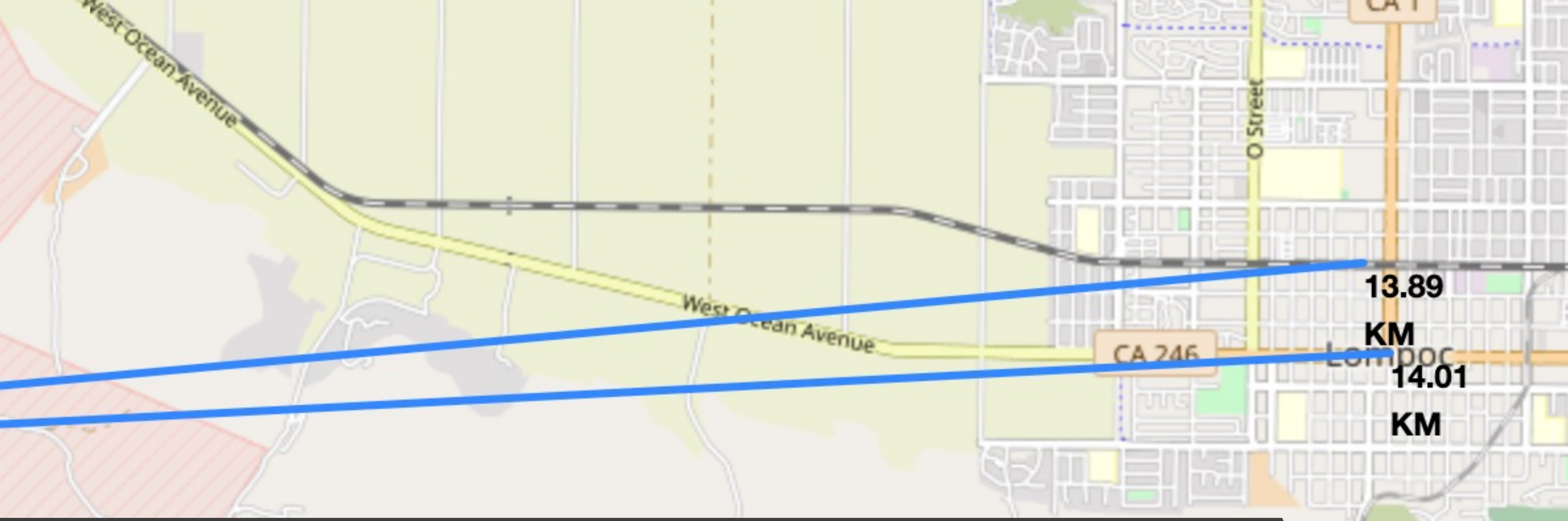
Launch Site to Landmark Distance

This image depicts the closest coastline, railroad, highway, and city to Vandenburg Air Force Base.



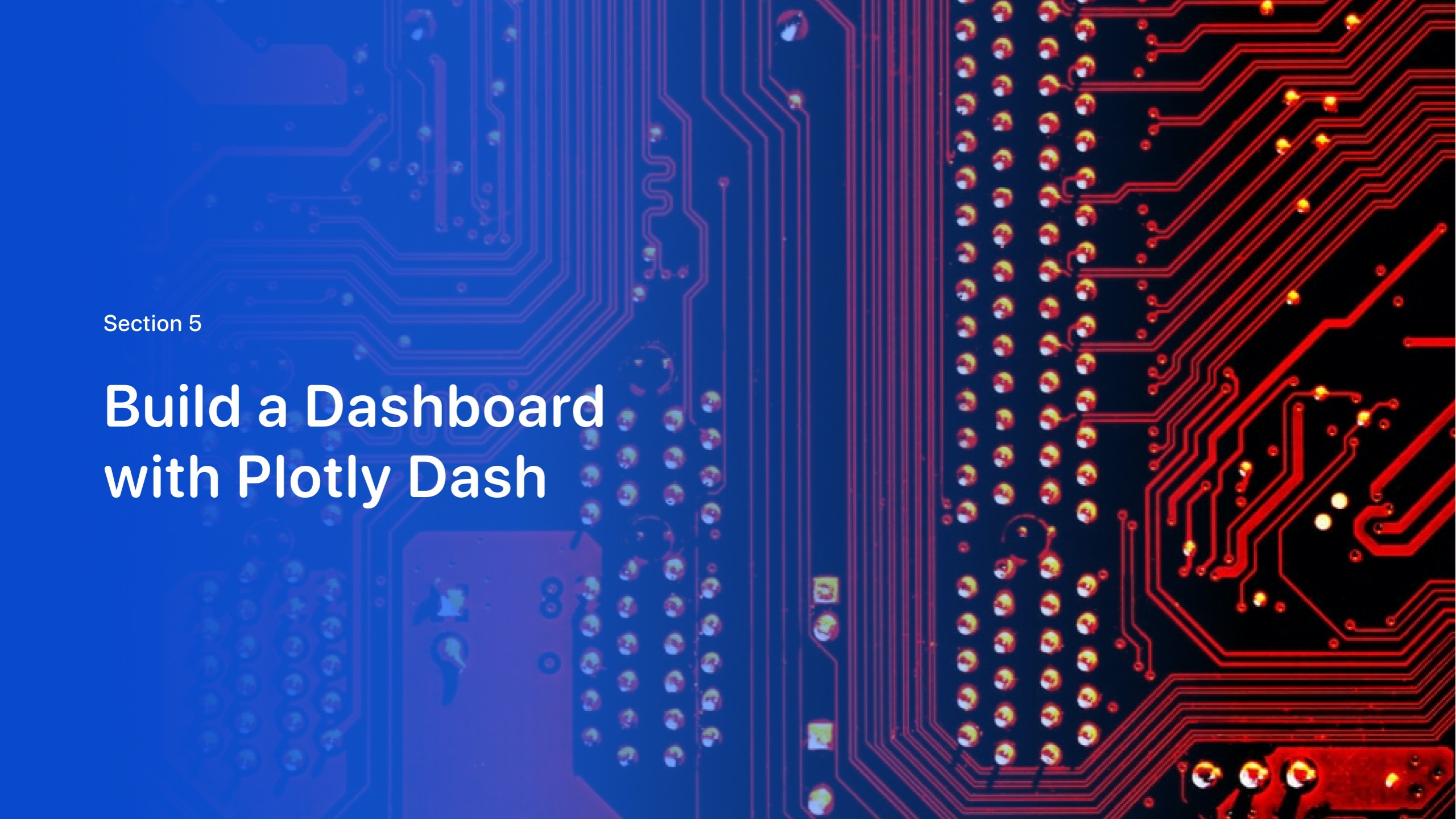
Closest Coastline and Railroad

- The closest coastline is 1.35 KM away
- The closets railroad is 1.26 KM away



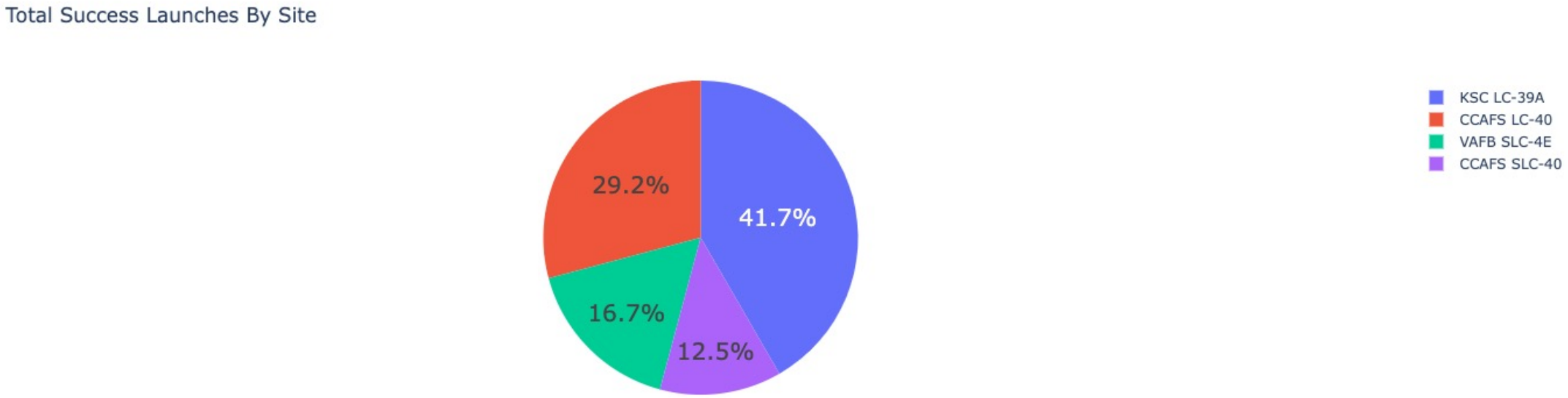
Closest City and Highway

- The closest city to VAFB is the city of Lompoc 13.89 KM away
- The closest highway is 14.01 KM away



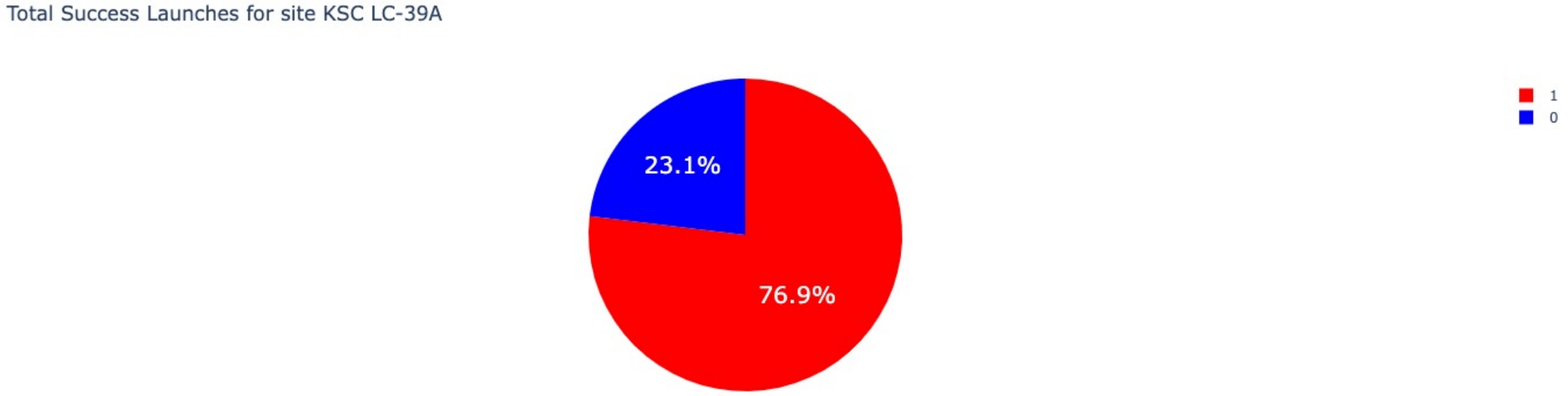
Section 5

Build a Dashboard with Plotly Dash



Visualizing Success Rates By Launch Site

- The Kennedy Space Center (KSC LC-39A) had the largest share of successful landings as seen in the pie chart.
- Cape Canaveral Air Force Station, however, does account for two of the launch sites CCAFS LC-40 and CCAFS SLC-40. Together these account for 41.7% of successful landings. On par with the Kennedy Space Center.



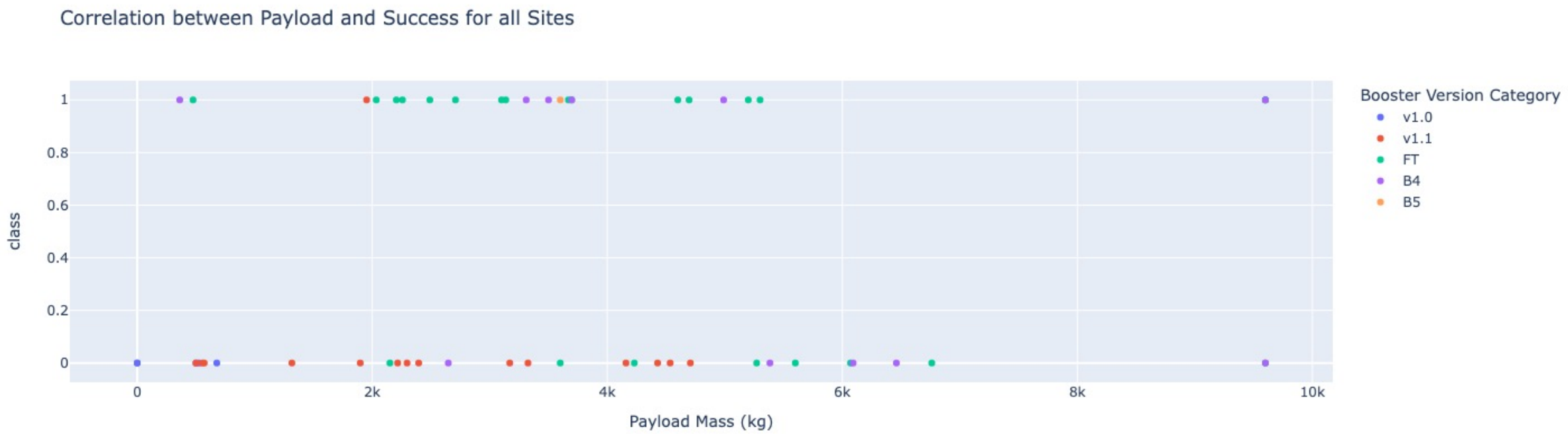
Success Rate at Kennedy Space Center Launch Site

- Successful landings (1) and unsuccessful landings (0) are depicted in red and blue, respectively.
- The Kennedy Space Center had the highest ratio of successful landings out of the four launch sites.

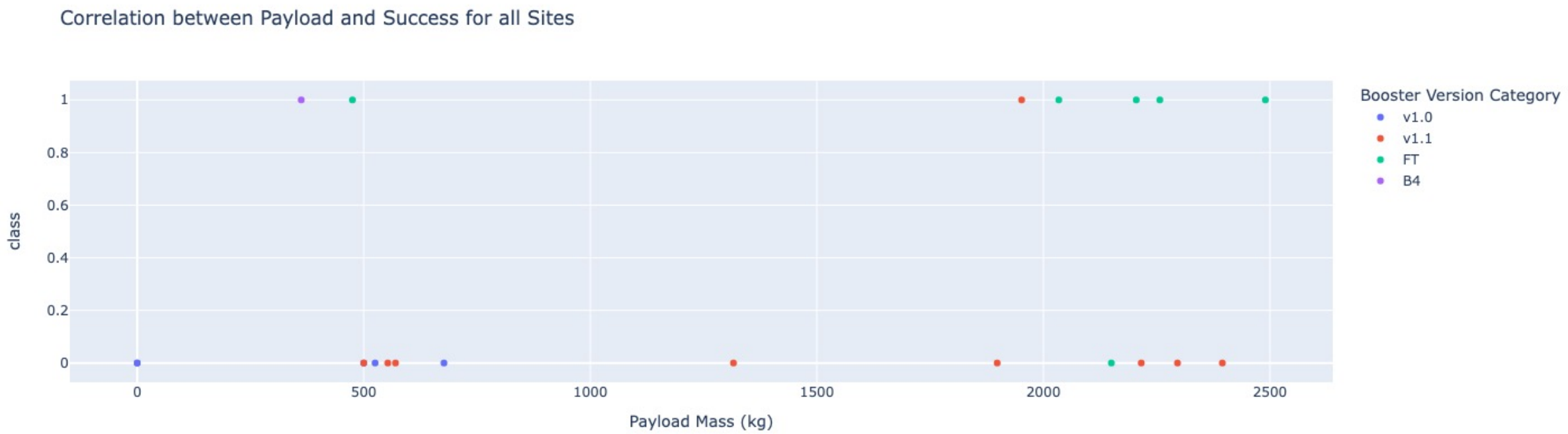
Payload Mass vs Landing Outcome

- The following four slides will depict four payload ranges, showing the effects of payload on the landing outcome.
- At the lower and higher ends of the range, there tends to be more failures.
- At a mid payload range (3K – 7K kg) there tends to be more success.

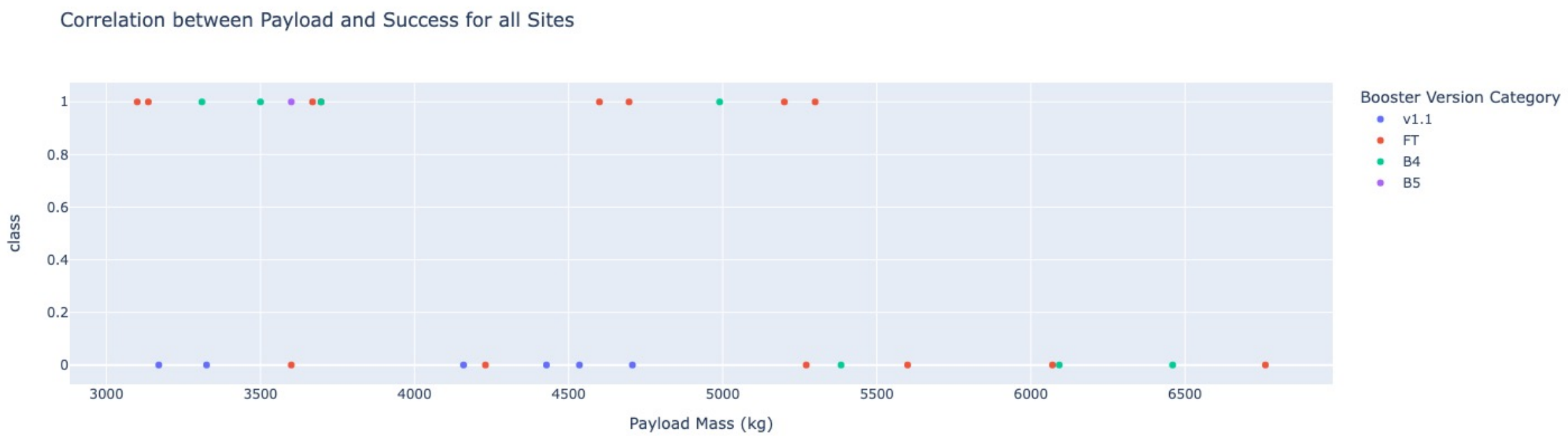
Full Payload Range (0 - 10K kg)



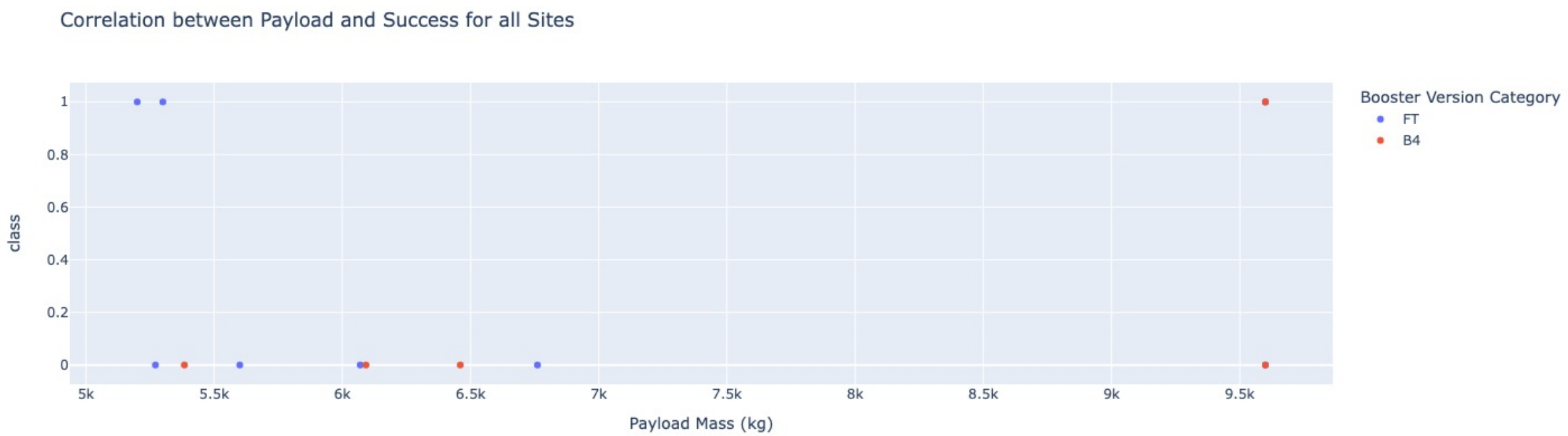
Low Payload Range (0 - 3K kg)



Mid Payload Range (3K— 7K kg)

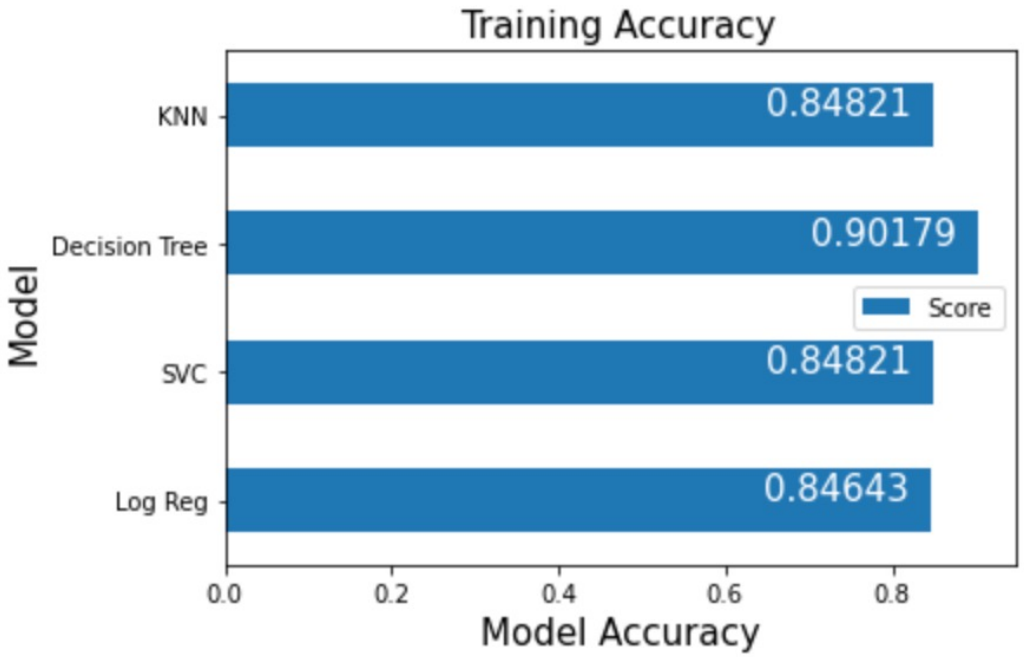


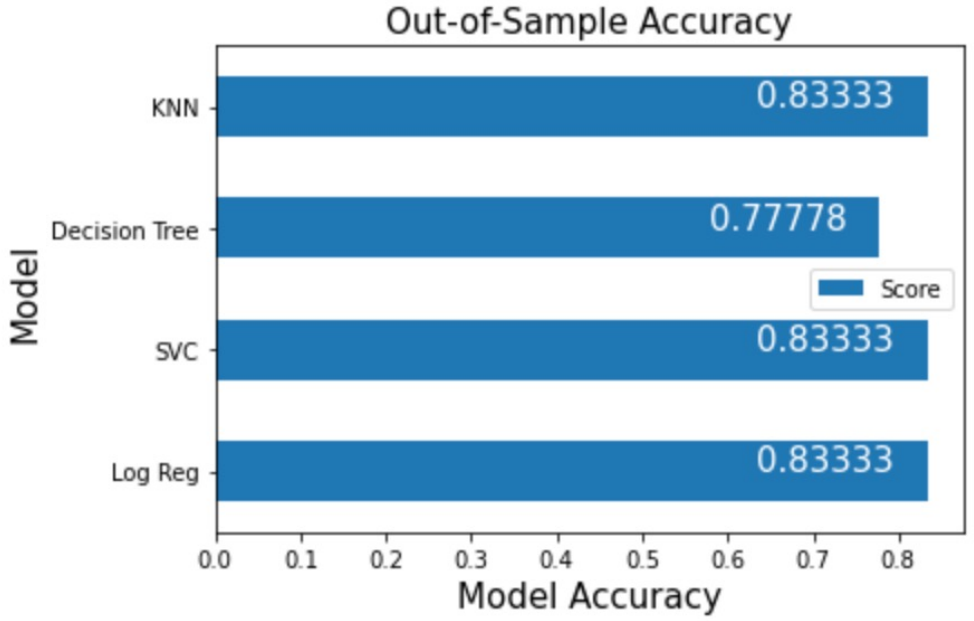
High Payload Range (5K – 10K kg)



Section 6

Predictive Analysis (Classification)



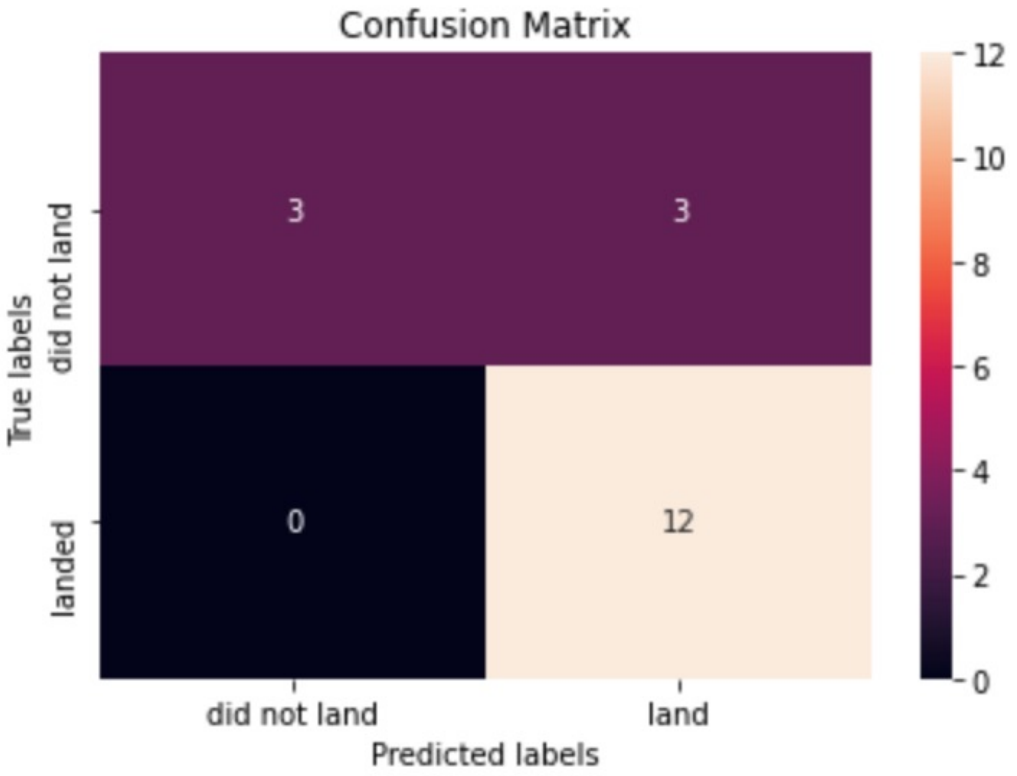


Classification Accuracy

- The Decision Tree had the highest training accuracy; However, it had the lowest out of sample accuracy.
- K Nearest Neighbors, SVC, and Logistic Regression have the highest out of sample accuracy.

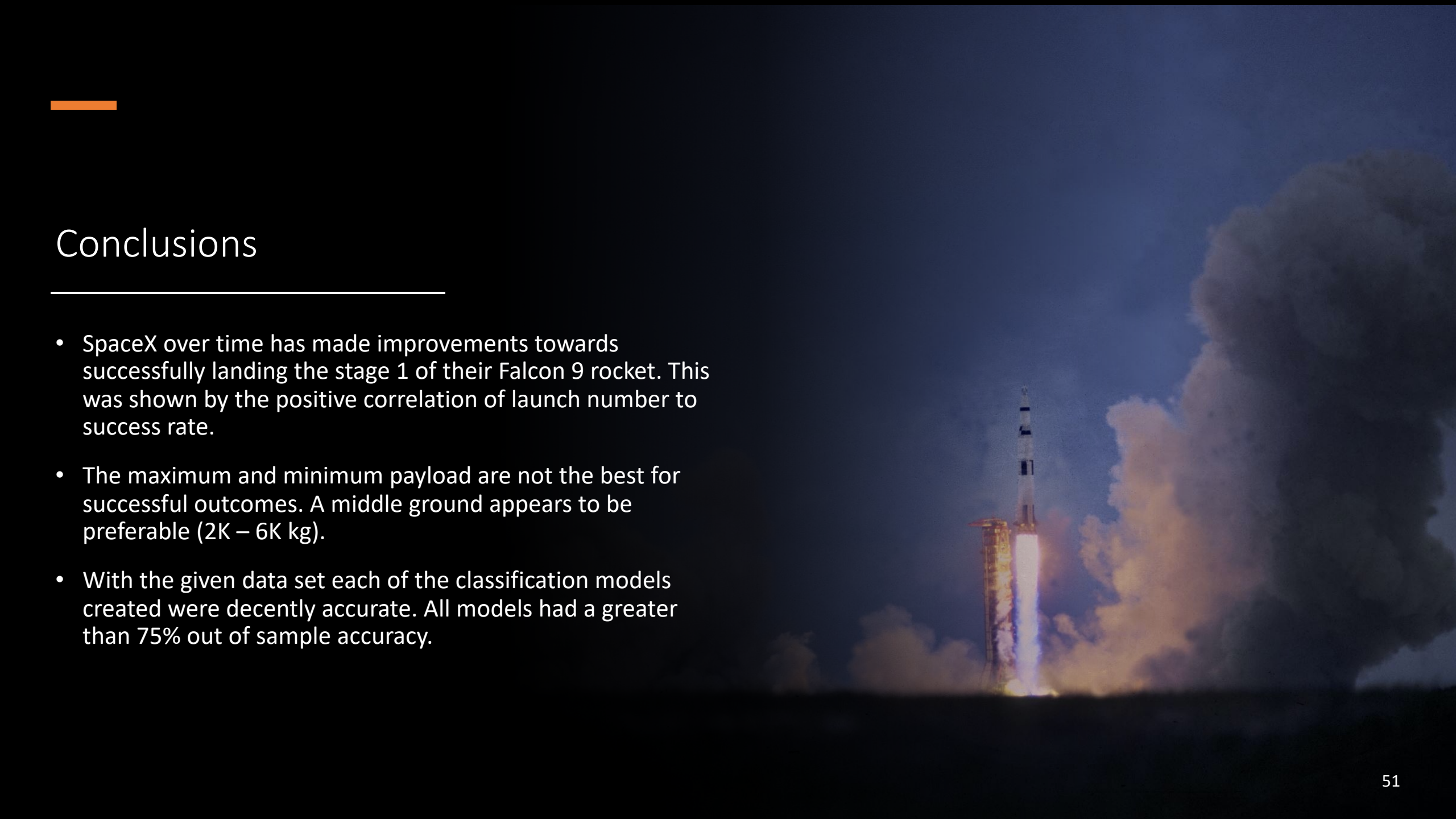
Confusion Matrix

- The three best performing models all produced the same confusion matrix.
- This matrix showed a 100% accuracy in detecting a successful landing.
- The accuracy was lower for unsuccessful landings, where 50% are correctly classified and 50% are false positives.



Conclusions

- SpaceX over time has made improvements towards successfully landing the stage 1 of their Falcon 9 rocket. This was shown by the positive correlation of launch number to success rate.
- The maximum and minimum payload are not the best for successful outcomes. A middle ground appears to be preferable (2K – 6K kg).
- With the given data set each of the classification models created were decently accurate. All models had a greater than 75% out of sample accuracy.



Thank you!