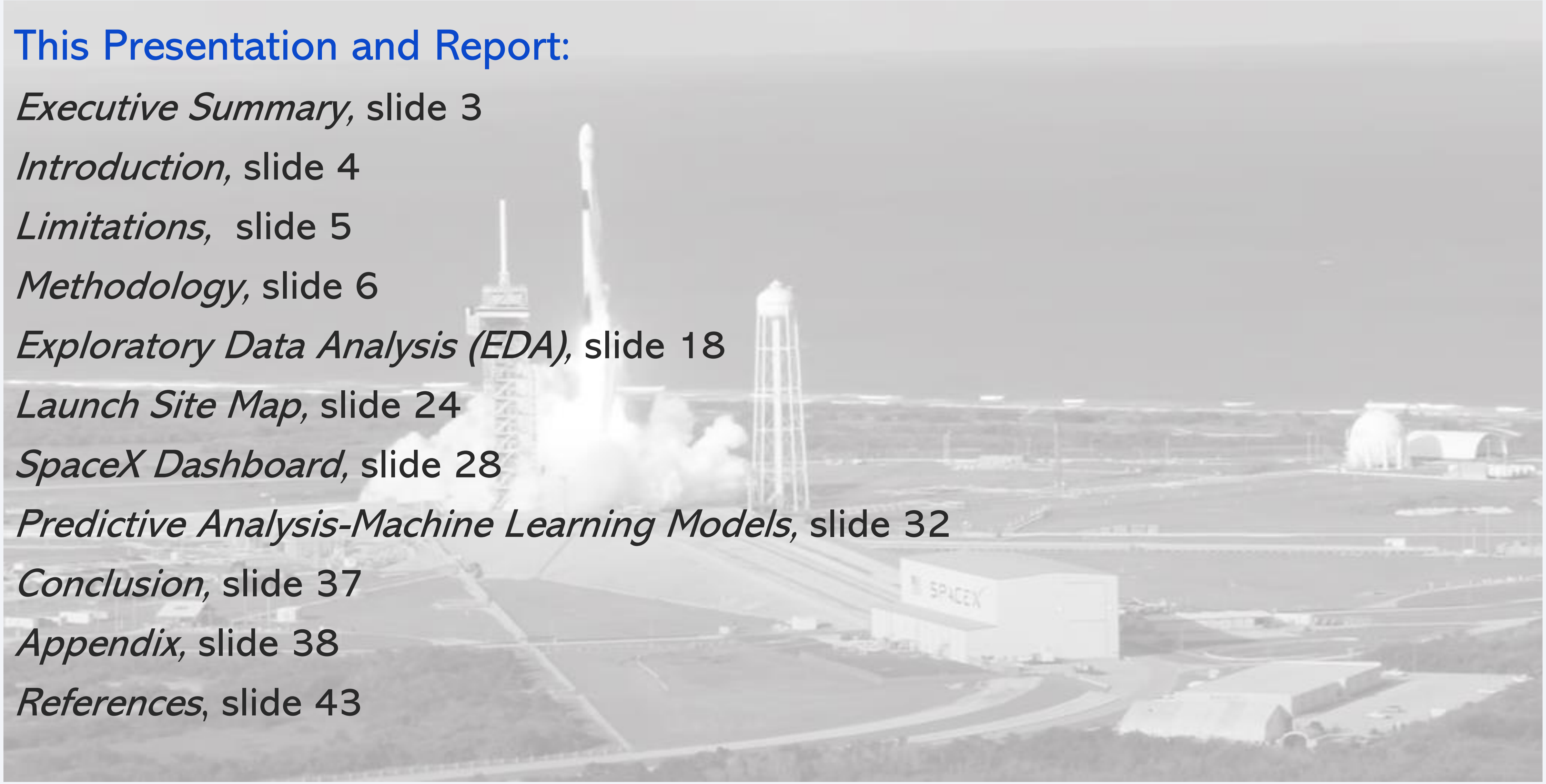
IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Dave Plumstead
August 11, 2024

Outline



This Presentation and Report:

Executive Summary, slide 3

Introduction, slide 4

Limitations, slide 5

Methodology, slide 6

Exploratory Data Analysis (EDA), slide 18

Launch Site Map, slide 24

SpaceX Dashboard, slide 28

Predictive Analysis-Machine Learning Models, slide 32

Conclusion, slide 37

Appendix, slide 38

References, slide 43

Executive Summary

The following presentation summarizes the SpaceX Falcon 9 project, which is the final assignment and capstone project in IBM’s Data Science program. The project incorporates many of the data science tools and techniques, and range of statistical tasks and work performed throughout the program.

Specifically, the project looks at the success rate of SpaceX Falcon 9 rocket booster (first stage) landings, and provides descriptive and inferential statistics to describe and predict the rocket landing outcomes.

The steps in the data and analytic process include collecting the SpaceX flight data; data preprocessing (data wrangling/ cleaning); exploratory data analysis and visualization; and statistical modeling with predictive machine learning classification models.

The key findings and results are summarized through the exploratory data analysis with descriptive charts and tables; an interactive SpaceX launch map and dashboard; and various predictive models which are developed and compared for performance.

Introduction

SpaceX has designed the Falcon 9 rocket to take people and other various payloads (including Starlink satellites) into ‘earth orbit and beyond’. The unique, two-stage rocket has first-stage booster parts that are reusable, making the falcon capable of ‘reflights’. This gives the falcon an apparent economic advantage over other, one-time (single-use) rockets, which lowers the cost of a rocket launch significantly and increases access to space.[1]

According to the SpaceX website and at the time of this report, the Falcon 9 rocket has made 359 launches, with 317 landings and 291 reflights (SpaceX, 2024).

Objective

In view of the above, the objective of the study is to analyze the success rate of Falcon 9 landings and develop a statistical model to predict the probability of successful landings for future flights. This provides insight into the cost of a SpaceX rocket launch based on whether or not the first stage booster can be reused. The statistical analysis and modeling can also be useful in other areas such as space industry research and helping to advance the space economy.

[1] For example, the SpaceX price chart has the cost of a Falcon 9 rocket at about $62M compared to other rocket providers of $165M.

Limitations

- The analysis is based on public SpaceX launch data for the period 2010-2020. The data could be updated to include more recent flights based on data available at the time of the study. This would also increase the sample size (see below).

- The data set for model prediction represents 90 SpaceX flights, which is a relatively small dataset that can result in partitioning (train/test) errors for some models such as decision trees (for example, an 80/20 train/test data split leaves just 18 outcome samples in the test set to evaluate model performance). Changing the model parameters to find the balance between model performance and accuracy while reducing errors on a small dataset, can be challenging.

- Depending on the analysis, different SpaceX datasets were used that had a different number of flight records and landing success rates (see also, Data Collection). For example, the datset used for the geo mapping and dashboard was a smaller dataset than the one used for the data exploration and statistical modeling. The data sets were not harmonized or joined which would provide a more cohesive and consistent data source.

- The REST API data includes other variables/ features that may be useful for developing predictive models, which were not included in the study. These include the *cost per launch* and *launch attempts*. Classification and model performance may be improved by exploring these other features and predictors of rocket landing outcomes.

- For the statistical models, the booster landing category ‘None None’ (see Appendix 3) is considered to be a negative outcome (i.e., unsuccessful booster, or first stage, landing). Depending on the interpretation of the landing definition (‘no attempt’) this category may not be considered an unsuccessful landing by some. As this landing category represents 21% of the Falcon 9 landing outcomes in the modeling data set, this would significantly change the results of the predictive models.

- The preliminary analysis and modeling in this study and report does not include domain expertise or input from SpaceX staff or space industry professionals. This would be recommended for any further research and statistical analysis.

Section 1

Methodology

Methodology- Summary

The SpaceX study and data collection and analysis was conducted using Python programming language. SQL was also used for some of the exploratory data analysis. Most of the coding and statistical work was performed in Jupyter Notebooks, with the Theia platform also used to create an interactive dashboard.

Data collection through a combination of REST APIs and web scraping.

Data pre-processing, cleaning, and wrangling.

Exploratory data analysis using SQL and data visualizations with Matplotlib and Seaborn Python libraries.

Interactive dashboard created using Folium and Plotly Dash.

Predictive analysis using Python's Scikit-learn library and machine learning classification models including Logistic Regression, Decision Trees, K Nearest Neighbor, and Support Vector Machine.

*Note: links to the various Jupyter Notebooks and code are provided in the relevant sections of the report.

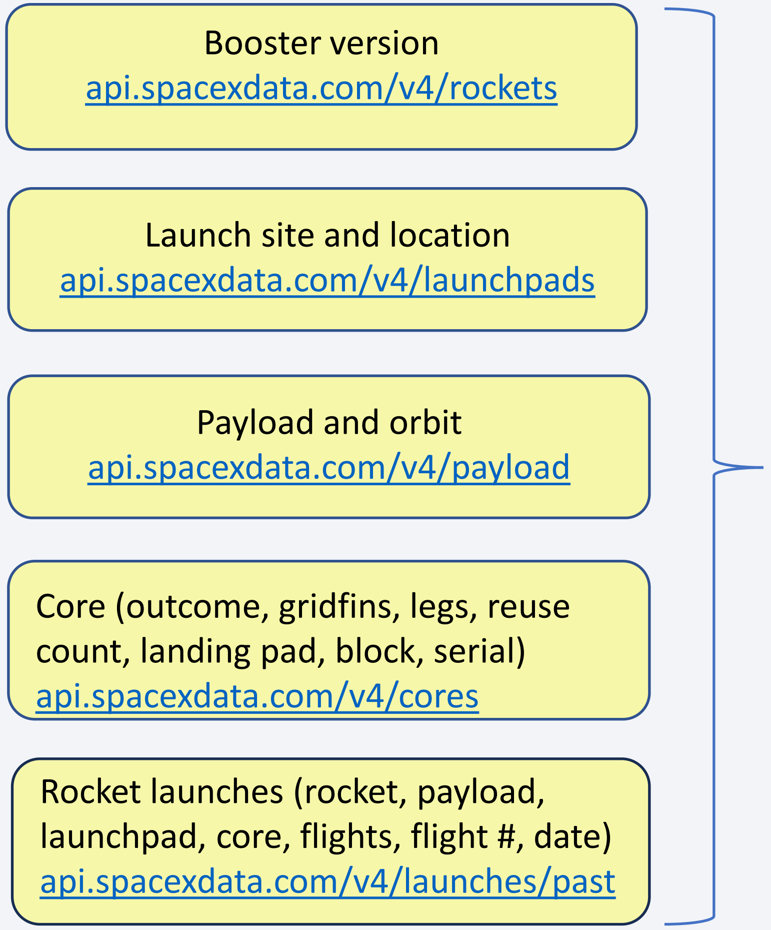
Data Collection – SpaceX REST API

- Open-source REST (Representational State Transfer) API (Application Programming Interface) used to collect SpaceX data, which includes flight, rocket, core, launch/ landing pad, outcome, and other data.
- Datasets extracted from the REST API url and various endpoints using Python *Requests* library, through an http GET request to the API.
- API response data in the form of JSON is transformed through Panda's *json normalize* function to put the data into a flat table (with columns).
- Final data set (17 variables) constructed by combining column data into a Python dictionary and then creating a Pandas data frame.
- Further data pre-processing, wrangling, and cleaning is required to format data for statistical analysis and modeling (see data preprocessing slide).

Jupyter Notebook GitHub link: [IBM-Data-Science/Lab1_jupyter-labs-spacex-data-collection-api.ipynb at main · DavePlum/IBM-Data-Science (github.com)]()

SpaceX REST API
https://api.spacexdata.com/v4/

Raw data:
SpaceX Dataset



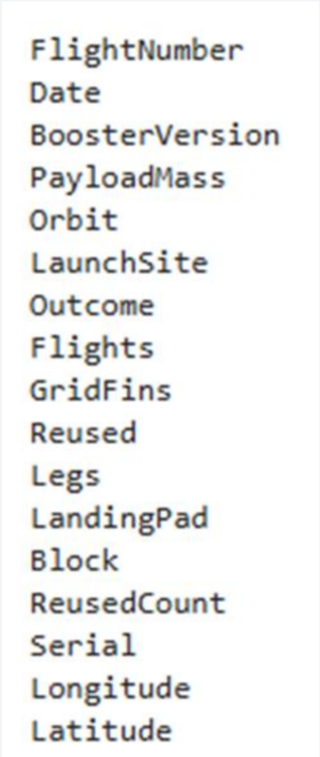
Booster version
api.spacexdata.com/v4/rockets

Launch site and location
api.spacexdata.com/v4/launchpads

Payload and orbit
api.spacexdata.com/v4/payload

Core (outcome, gridfins, legs, reuse count, landing pad, block, serial)
api.spacexdata.com/v4/cores

Rocket launches (rocket, payload, launchpad, core, flights, flight #, date)
api.spacexdata.com/v4/launches/past

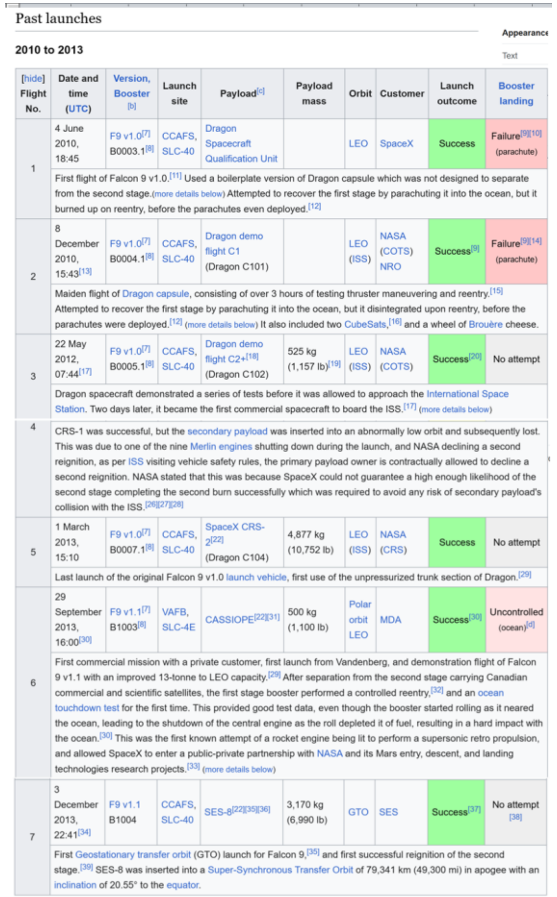


```
FlightNumber
Date
BoosterVersion
PayloadMass
Orbit
LaunchSite
Outcome
Flights
GridFins
Reused
Legs
LandingPad
Block
ReusedCount
Serial
Longitude
Latitude
```

Data Collection – SpaceX Web Scraping

- SpaceX data was also collected by scraping flight record data from tables on Wikipedia: List of Falcon 9 and Falcon Heavy launches – Wikipedia (June, 2021 revision).
- The html tables contain similar variables to the previous data collection (REST API) and also new ones including flight customer, payload, and launch and landing outcomes.
- Data scraped from 'Past launches' tables (2010- 2021) using Python *Beautiful soup* and *Request* libraries.
- Html tables parsed by extracting html column/variable names, creating a Python dictionary, and converting the dictionary to a Pandas data frame.
- Final scraped dataset (227 records/ flights; 11 variables) exported to a csv file.

Jupyter Notebook GitHub link: IBM-Data-Science/Lab2_jupyter-labs-webscraping.ipynb at main · DavePlum/IBM-Data-Science (github.com)



Past launches

Appearance

Text

2010 to 2013

[hide] Flight No.	Date and time (UTC)	Version, Booster [b]	Launch site	Payload[c]	Payload mass	Orbit	Customer	Launch outcome	Booster landing
1	4 June 2010, 18:45	F9 v1.0[7] B0003.1[8]	CCAFS, SLC-40	Dragon Spacecraft Qualification Unit		LEO	SpaceX	Success	Failure[9][10] (parachute)
	First flight of Falcon 9 v1.0.[11] Used a boilerplate version of Dragon capsule which was not designed to separate from the second stage.(more details below) Attempted to recover the first stage by parachuting it into the ocean, but it burned up on reentry, before the parachutes even deployed.[12]								
2	8 December 2010, 15:43[13]	F9 v1.0[7] B0004.1[8]	CCAFS, SLC-40	Dragon demo flight C1 (Dragon C101)		LEO (ISS)	NASA (COTS) NRO	Success[9]	Failure[9][14] (parachute)
	Maiden flight of Dragon capsule, consisting of over 3 hours of testing thruster maneuvering and reentry.[15] Attempted to recover the first stage by parachuting it into the ocean, but it disintegrated upon reentry, before the parachutes were deployed.[12] (more details below) It also included two CubeSats,[16] and a wheel of Brouère cheese.								
3	22 May 2012, 07:44[17]	F9 v1.0[7] B0005.1[8]	CCAFS, SLC-40	Dragon demo flight C2+[18] (Dragon C102)	525 kg (1,157 lb)[19]	LEO (ISS)	NASA (COTS)	Success[20]	No attempt
	Dragon spacecraft demonstrated a series of tests before it was allowed to approach the International Space Station. Two days later, it became the first commercial spacecraft to board the ISS.[17] (more details below)								
4	CRS-1 was successful, but the secondary payload was inserted into an abnormally low orbit and subsequently lost. This was due to one of the nine Merlin engines shutting down during the launch, and NASA declining a second reignition, as per ISS visiting vehicle safety rules, the primary payload owner is contractually allowed to decline a second reignition. NASA stated that this was because SpaceX could not guarantee a high enough likelihood of the second stage completing the second burn successfully which was required to avoid any risk of secondary payload's collision with the ISS.[26][27][28]								
5	1 March 2013, 15:10	F9 v1.0[7] B0007.1[8]	CCAFS, SLC-40	SpaceX CRS-2[22] (Dragon C104)	4,877 kg (10,752 lb)	LEO (ISS)	NASA (CRS)	Success	No attempt
	Last launch of the original Falcon 9 v1.0 launch vehicle, first use of the unpressurized trunk section of Dragon.[29]								
6	29 September 2013, 16:00[30]	F9 v1.1[7] B1003[8]	VAFB, SLC-4E	CASSIOPE[22][31]	500 kg (1,100 lb)	Polar orbit LEO	MDA	Success[30]	Uncontrolled (ocean)[d]
	First commercial mission with a private customer, first launch from Vandenberg, and demonstration flight of Falcon 9 v1.1 with an improved 13-tonne to LEO capacity.[29] After separation from the second stage carrying Canadian commercial and scientific satellites, the first stage booster performed a controlled reentry,[32] and an ocean touchdown test for the first time. This provided good test data, even though the booster started rolling as it neared the ocean, leading to the shutdown of the central engine as the roll depleted it of fuel, resulting in a hard impact with the ocean.[30] This was the first known attempt of a rocket engine being lit to perform a supersonic retro propulsion, and allowed SpaceX to enter a public-private partnership with NASA and its Mars entry, descent, and landing technologies research projects.[33] (more details below)								
7	3 December 2013, 22:41[34]	F9 v1.1 B1004	CCAFS, SLC-40	SES-8[22][35][36]	3,170 kg (6,990 lb)	GTO	SES	Success[37]	No attempt [38]
	First Geostationary transfer orbit (GTO) launch for Falcon 9,[35] and first successful reignition of the second stage.[39] SES-8 was inserted into a Super-Synchronous Transfer Orbit of 79,341 km (49,300 mi) in apogee with an inclination of 20.55° to the equator.								

Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
1	CCAFS	Dragon Space	0	LEO	SpaceX	Success	F9 v1.0B0003.1	Failure	04-Jun-10	18:45
2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0004.1	Failure	08-Dec-10	15:43
3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0005.1	No attempt	22-May-12	7:44
4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success	F9 v1.0B0006.1	No attempt	08-Oct-12	0:35
5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success	F9 v1.0B0007.1	No attempt	01-Mar-13	15:10
6	VAFB	CASSIOPE	500 kg	Polar orbit	MDA	Success	F9 v1.1B1003	Uncontrolled	29-Sep-13	16:00
7	CCAFS	SES-8	3,170 kg	GTO	SES	Success	F9 v1.1	No attempt	03-Dec-13	22:41

Data Preprocessing (Wrangling and Cleaning)

In addition to the data preprocessing described above under data collection, the following summarizes further data wrangling and cleaning necessary to prepare the dataset for analysis and the predictive classification models:

- Data filtered to only include Falcon 9 flights and the flight numbers were reset.
- Missing values for 'Payload' were replaced with the average payload.
- Blank rows were removed from the dataset.
- From the exploratory data analysis 12 variables/features were identified as having the strongest influence on Falcon 9 landing outcomes. A new data frame (x) was created with these features for the predictive modeling.
- The booster landing feature has eight categories (see Appendix 3). Based on a roll-up and combination of these categories, a new feature *Class* was created where a successful landing = 1 and an unsuccessful landing = 0. This is the target/outcome variable (y) for prediction in the machine learning classification models.
- For the machine learning models, the categorical variables were changed to binary numerical values using one hot encoding (Pandas *get_dummies())*.
- The numerical variables were changed from integers to float.
- The 'Class' variable (above) was converted from a column data table to a NumPy array and assigned to the variable y in the models.
- The features in the dataset (x) were standardized and reassigned to the variable x (as a NumPy array) in the models using StandardScaler from scikit-learn's preprocessing module.

Jupyter Notebook GitHub link: IBM-Data-Science/Lab3_jupyter-spacex-Data wrangling.ipynb at main · DavePlum/IBM-Data-Science (github.com)

Exploratory Data Analysis with Visualization

After collecting and preprocessing the data described earlier, the next step in the analytics process is exploring the data to look at launch variables and features of interest, and to see what the data is saying.

Data exploration involves summary descriptive statistics with frequency distributions and tables, and looking at patterns, relationships and trends in the data through visualizations such as scatter, bar, and line charts.

Using Python's *Pandas*, *NumPy*, *Seaborn*, and *Matplotlib* libraries the data exploration included:

- Examining the influence of the flight number and payload mass (kg) on the first stage landing outcome.
- Looking at the relationships between landing outcome and payload mass for different Falcon 9 launch sites.
- Examining the relationship between flight number, landing success rate, and payload with the type of orbit (see Appendix 4).
- Analyzing the landing success rate trend during the 10-year period.
- Determining which features in the dataset are the best predictors of a Falcon 9 landing outcome, for predictive modeling.

Jupyter Notebook GitHub link: [IBM-Data-Science/Lab5_jupyter-spacex-edadataviz.ipynb at main · DavePlum/IBM-Data-Science (github.com)]

Exploratory Data Analysis with SQL

For further data analysis, SQL (Structured Query Language) was integrated with the Python environment using *ipython-sql* and *sqlalchemy* packages, and executed in Jupyter Notebook using SQL Magic commands.

This enabled SQL to be used to query the SpaceX database and answer key questions such as those in the table below (see EDA slide # 22-23 for query code and results).

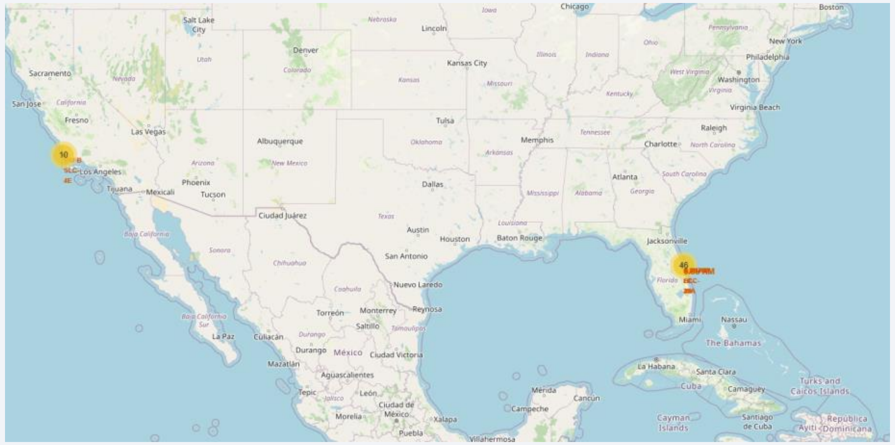
What are the names of the Falcon 9 launch sites? Show the first five records in the database where the launch site is in Florida (i.e., name begins with 'CCA').
What is the total payload mass, for Falcon 9 flights carried out for NASA (CRS)?
What is the average payload mass for a particular booster version (F9 v1.1)?
When was the first successful landing outcome on the ground pad achieved?
Which booster versions have had successful landings on a drone ship, with payloads ranging between 4,000-6,000 kg?
What is the number of successful and failed mission outcomes?
Which booster versions have carried the maximum payload mass?
In 2015, what are the dates, booster version, and launch site where there was an unsuccessful landing on the drone ship?
What are the various landing outcomes between June 4, 2010, and March 20, 2017?

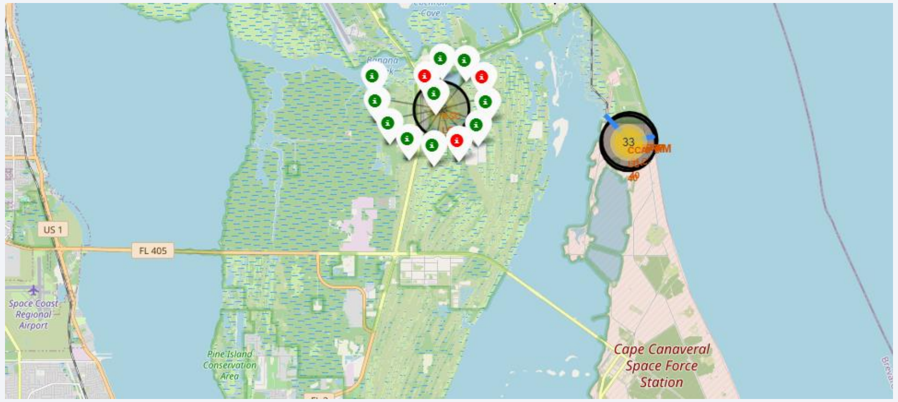
Jupyter Notebook GitHub link: IBM-Data-Science/Lab4_jupyter-labs-eda-sql-coursera_sqllite.ipynb at main · DavePlum/IBM-Data-Science (github.com)

Launch Site Map and Analysis

The study methodology also included developing an interactive map for geographic and visual analysis of the Falcon 9 launch sites. The map shows the location of the launch sites and the associated landing success rates, and proximities to areas of interest.

- Map developed using Python library *Folium* and various plugins (marker cluster, mouse position, features, poly line).
- Launch sites shown on map using lat and long coordinates, and marked by circles and name of site.
- Landing outcome /success rate for each launch pad flight shown using Folium color marker clusters (green=success, red=fail).
- Distance from launch pad to nearest coastline and railways calculated using python *math* module and shown with a straight line.




33
US 1
FL 405
Space Coast Regional Airport
Pine Island Conservation Area
Cape Canaveral Space Force Station

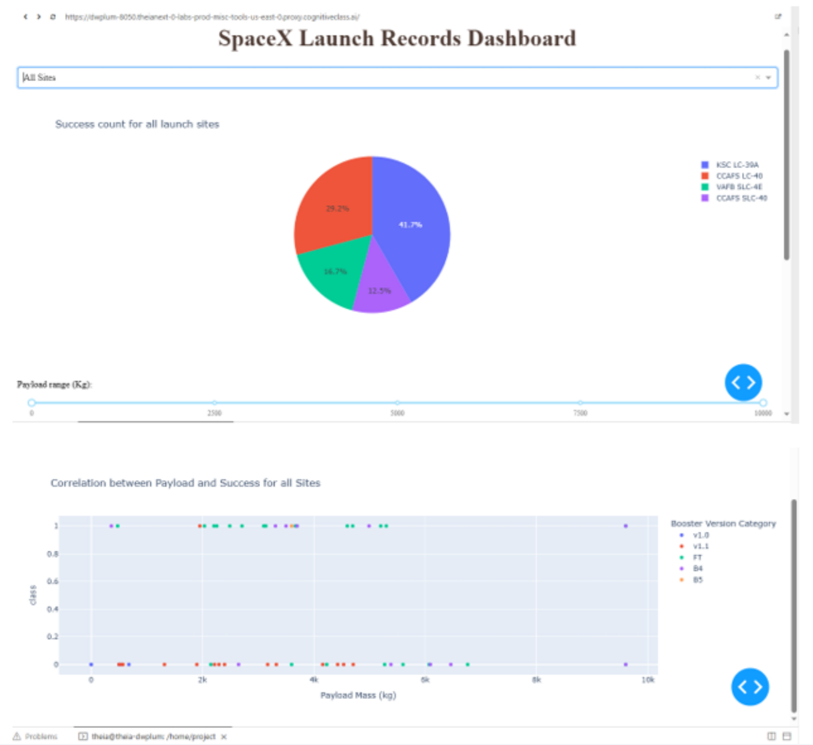
13

Jupyter Notebook GitHub link: [IBM-Data-Science/Lab6_jupyter_launch_site_location.ipynb at main · DavePlum/IBM-Data-Science (github.com)]

SpaceX Interactive Dashboard

An interactive dashboard was developed to provide additional visual analytics and insights from the SpaceX dataset. Specifically, the dashboard can be filtered by launch site to show the associated number of successful landings and the success rate for each site. The dashboard also explores the relationship between the landing success rate, payload mass, and launch sites.

- Dashboard developed using Python *Plotly Dash.* The dashboard was developed on the Theia platform and then adapted for Jupyter Notebook and Google Colab for user viewing (see Github links below. Colab link on slides 29-31).
- Skelton dash app used for initial coding and dashboard functionality and layout.
- Coding added for drop-down menu and callback function, to show successful launch counts and rates in pie chart.
- Coding added for payload range slider and callback function, to show relationship between landing success rate and payload (kg) in scatter plot.



https://dwplum-8050.theianext-0-labs-prod-misc-tools-us-east-0.proxy.cognitiveclass.ai/

SpaceX Launch Records Dashboard

All Sites

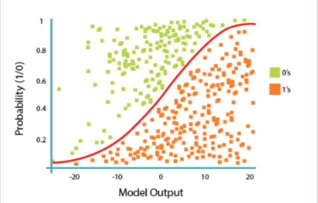
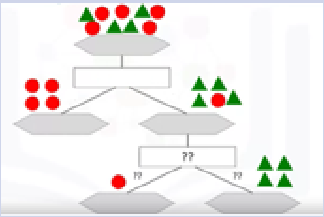
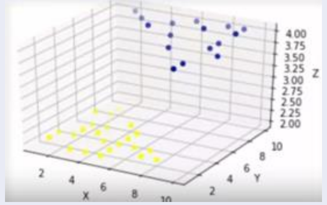
Payload range (Kg):

Jupyter Notebook GitHub link: IBM-Data-Science/Lab7_SpaceX_dashboard.ipynb at main · DavePlum/IBM-Data-Science (github.com)

Python code (Theia IDE) GitHub link: IBM-Data-Science/Lab7_SpaceX_dashboard_appcode_Theia.py at main · DavePlum/IBM-Data-Science (github.com)

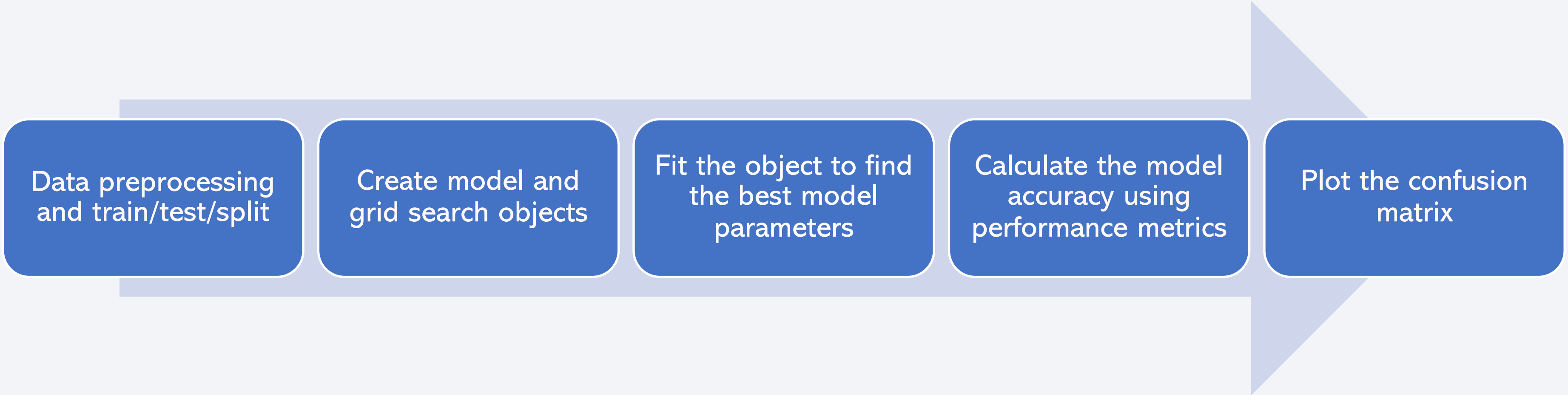
Predictive Analysis-Machine Learning Models

To predict the outcome of a Falcon 9 rocket landing, the following machine learning classification models were developed and compared in predictive performance:

Classification Model	Method	Model equations	Data distribution
Logistic regression	Transform data with an exponential function to predict the class (successful/unsuccessful) of a falcon 9 landing and the probability of the landing outcome.	Logistic function Logit(pi) = 1/(1+ exp(-pi)) ln(pi/(1-pi)) = Beta_0 + Beta_1*X_1 + … + B_k*K_k	
K-Nearest Neighbor	Classify cases based on similarity to other cases to predict the class (successful or unsuccessful landing) of a Falcon 9 flight.	Euclidean distance $d(x,y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$	1. Basic Service 2. E-Service 3. Plus Service 4. Total Service ?
Decision Trees	Split the dataset into subsets using strongest predictors of outcome (success or failure) and through recursive partitioning, optimize the model to predict – or decide – whether a Falcon 9 rocket will land successfully or not.	Entropy = - p(A)$\log_2$(p(A)) - p(B)$\log_2$(p(B)) Information Gain = (entropy before split) – (weighted entropy after split)	
Support Vector Machine	Transform data to a higher-dimensional space and find a hyperplane that best divides the data set into two classes (successful or unsuccessful landing).	Depends on the *kernel* function (linear, polynomial, radial basis, sigmoid)	

Predictive Analysis - Model Development Process

The diagram below shows the general steps for developing the SpaceX classification models (see Appendix 5 for model parameters):



Data preprocessing and train/test/split

Create model and grid search objects

Fit the object to find the best model parameters

Calculate the model accuracy using performance metrics

Plot the confusion matrix

Jupyter Notebook GitHub link: [IBM-Data-Science/Lab8_SpaceX_Machine Learning Prediction.ipynb at main · DavePlum/IBM-Data-Science (github.com)]

Predictive Analysis – Model Development and Evaluation

The classification models (previous slide) were developed using a number of Python libraries including *scikit-learn* for machine learning model development.

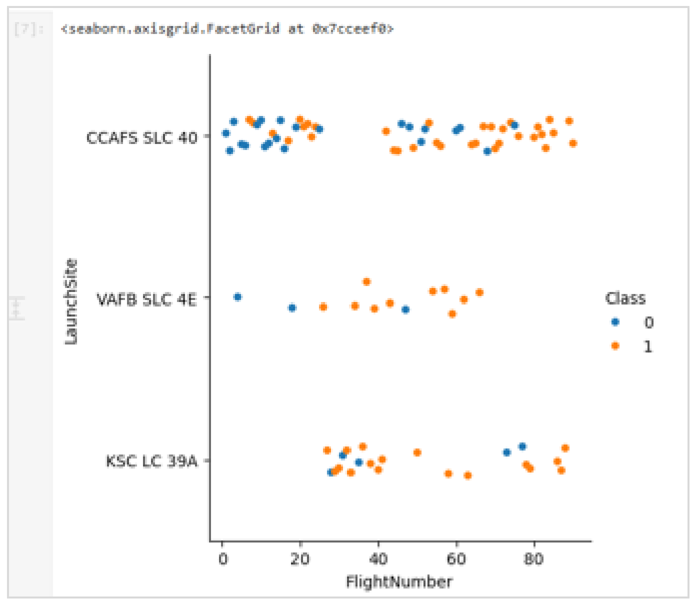
- Data transformation included converting the *Class* (outcome variable) to a *NumPy array* (y) and standardizing the features/predictors (x) so they have similar scales for the models (see also, Data Preprocessing).
- The SpaceX data features (x) and outcome *Class* (y) were split into training and testing data using *sklearn's train_test_split* function (80/20 split). The *random state* parameter was set at 2 to ensure a consistent train/test split each time the model was run and for model comparability.
- The performance of each model was optimized through *hyperparameter tuning* using Scikit learn's *cross-validation* and *grid search* method.
- The grid search splitting of the train/test data was set at a 10-fold cross validation (cv).
- The models predictive performance was evaluated using *sklearn's accuracy score* and *Confusion Matrix* (true/false positives; true/false negatives- see slide 34).

Section 2

Insights drawn from EDA

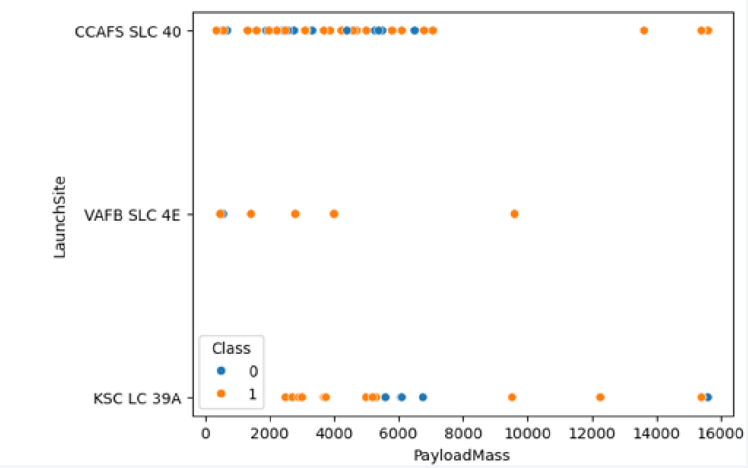
Relationship between launch site, flight number, payload

Class
● 0 Unsuccessful landing
● 1 Successful landing



```
<seaborn.axisgrid.FacetGrid at 0x7cceef0>
```

- Overall, the Falcon 9 has a landing success rate of 67% during the period. As noted from the chart, the rate varies by launch site.[2]
- During the period, site CCAFS SLC 40 launched the most flights (61%) and had a landing success rate of 60%.
- The other two sites (VAFB SLC 4E and KSC LC 39A) launched fewer flights (14.5% and 24.5% respectively) over the same period but had a higher success rate of 77%.
- The launch sites have a relatively small influence on rocket landing outcomes (the landing pads have a greater influence-see slide 36).
- The three sites appear to be trending towards higher landing success rates as the flight number increases. [3]

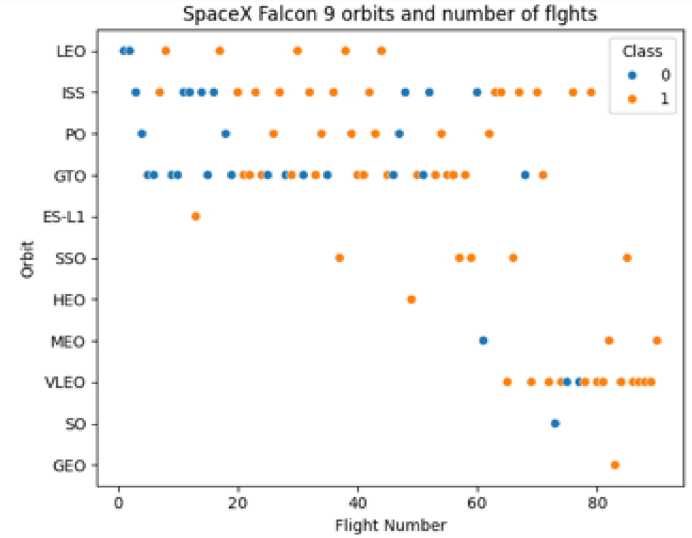


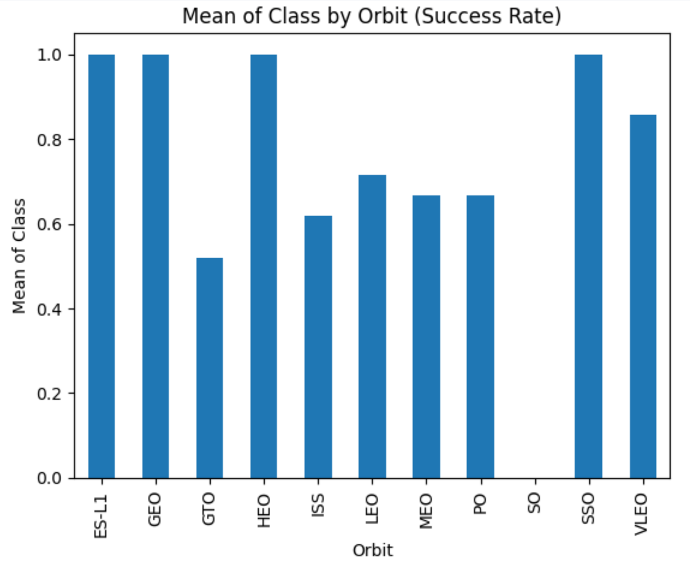
- The relationship between landing outcomes, and launch sites and rocket payload (kgs) is not as clear, as successful/unsuccessful landings occur for both, lighter and heavier payloads.
- It can be noted that rockets launching from site VAFB SLC 4E did not carry payloads over 10,000 kg. during the period.
- While it appears that heavier payloads may increase the likelihood of successful landings, the statistical modeling indicates that payload has a relatively small effect on the landing outcome.

[2] The *landing success rate* is calculated as the mean of the binary landing outcome, 'Class' (0 or 1).

[3] The 'Flight Number' corresponds with consecutive flight dates and thus can be viewed as a running total of the number of flights during the period (rather than a unique flight ID with other meaning).

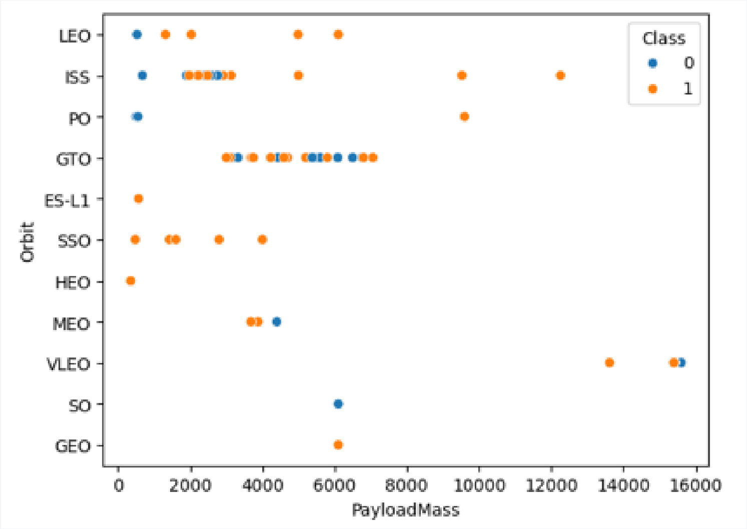
Orbit type and flight number, landing success rate



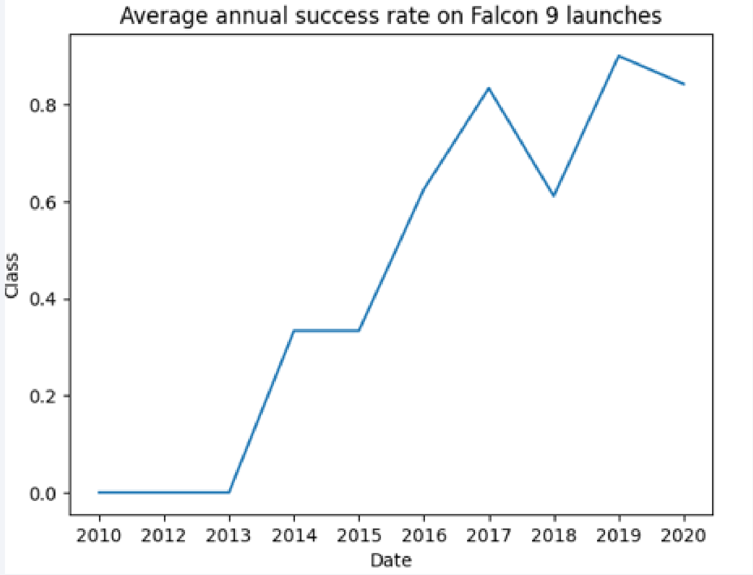


- During the period, SpaceX launched the Falcon 9 rocket into 11 different orbits (see Appendix 4 for orbit descriptions).
- A little over half (53%) the flights were flown to the International Space Station (ISS) and geosynchronous orbit (GTO).
- Only one flight was flown to each of the ES-L1, GEO, HEO, and SO orbits.
- The remaining orbits had a range of between 3 to 14 flights.
- In view of the above, caution needs to be applied when viewing the landing success rates by orbit type.
- For example, rockets returning from the ES-L1, HEO, and GEO orbits have a 100% landing success rate but this is based on one flight. On the flip side, the SO orbit has a 0% success rate but also based on one flight.
- It's interesting to note that flights to the sun-synchronous orbit (SSO) have a perfect landing success rate out of a handful of flights.
- The Very Low Earth Orbits (VLEO) also have a relatively high landing success rate (85.5%)
- For the more popular orbits, the Falcon 9 has a 62% landing success rate on flights to the International Space Station and a 52% rate on geosynchronous orbit flights.
- The statistical modeling indicates that orbits have a notable impact on landing outcomes.

Orbit type and payload, success rate trend

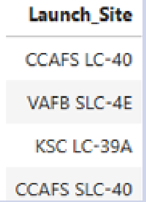
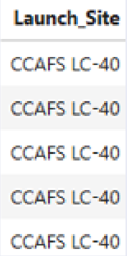
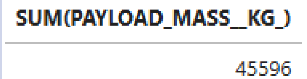
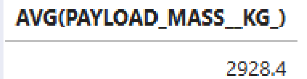
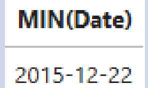
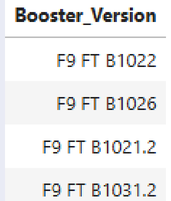


- Flights with heavier payloads appear to have successful landings returning from the polar orbit (PO) and International Space Station (ISS).
- The relationship is not as clear for the other orbits.

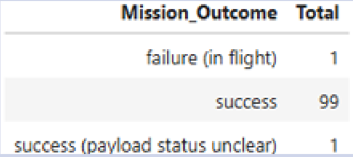
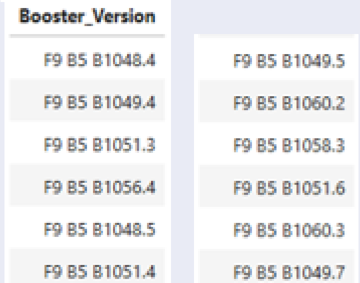
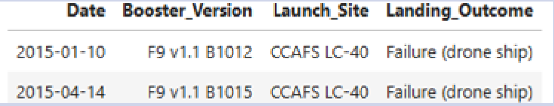
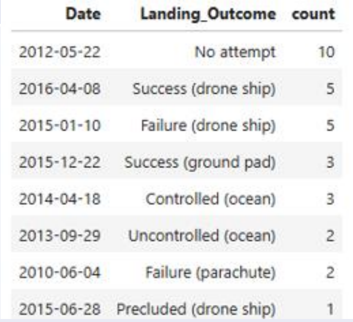


- Turning to trends, the Falcon 9 average landing success rate has been increasing over the 10-year period, apart from a few isolated years and down-turns.

Insights from SQL queries

Data exploration	SQL Query	Query Result
Unique names of the Falcon 9 launch sites.	%sql select DISTINCT ("launch_Site") from SPACEXTABLE	 Launch_Site CCAFS LC-40 VAFB SLC-4E KSC LC-39A CCAFS SLC-40
The first five records in the database where the launch site is in Florida (i.e., name begins with 'CCA').	%sql select Launch_Site from SPACEXTABLE WHERE Launch_Site LIKE '%CCA%' LIMIT 5	 Launch_Site CCAFS LC-40 CCAFS LC-40 CCAFS LC-40 CCAFS LC-40 CCAFS LC-40
The total payload mass for Falcon 9 flights carried out for NASA (CRS).	%sql select SUM(PAYLOAD_MASS__KG_) from SPACEXTABLE WHERE Customer = 'NASA (CRS)'	 SUM(PAYLOAD_MASS__KG_) 45596
Average payload mass for booster version F9 v1.1.	%sql select AVG(PAYLOAD_MASS__KG_) from SPACEXTABLE WHERE Booster_Version = 'F9 v1.1'	 AVG(PAYLOAD_MASS__KG_) 2928.4
Date of first successful landing outcome on the ground pad.	%sql select MIN(Date) from SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)'	 MIN(Date) 2015-12-22
Names of boosters that have had successful landings on a drone ship, with payloads ranging between 4,000-6,000 kg.	%sql select Booster_Version from SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000	 Booster_Version F9 FT B1022 F9 FT B1026 F9 FT B1021.2 F9 FT B1031.2

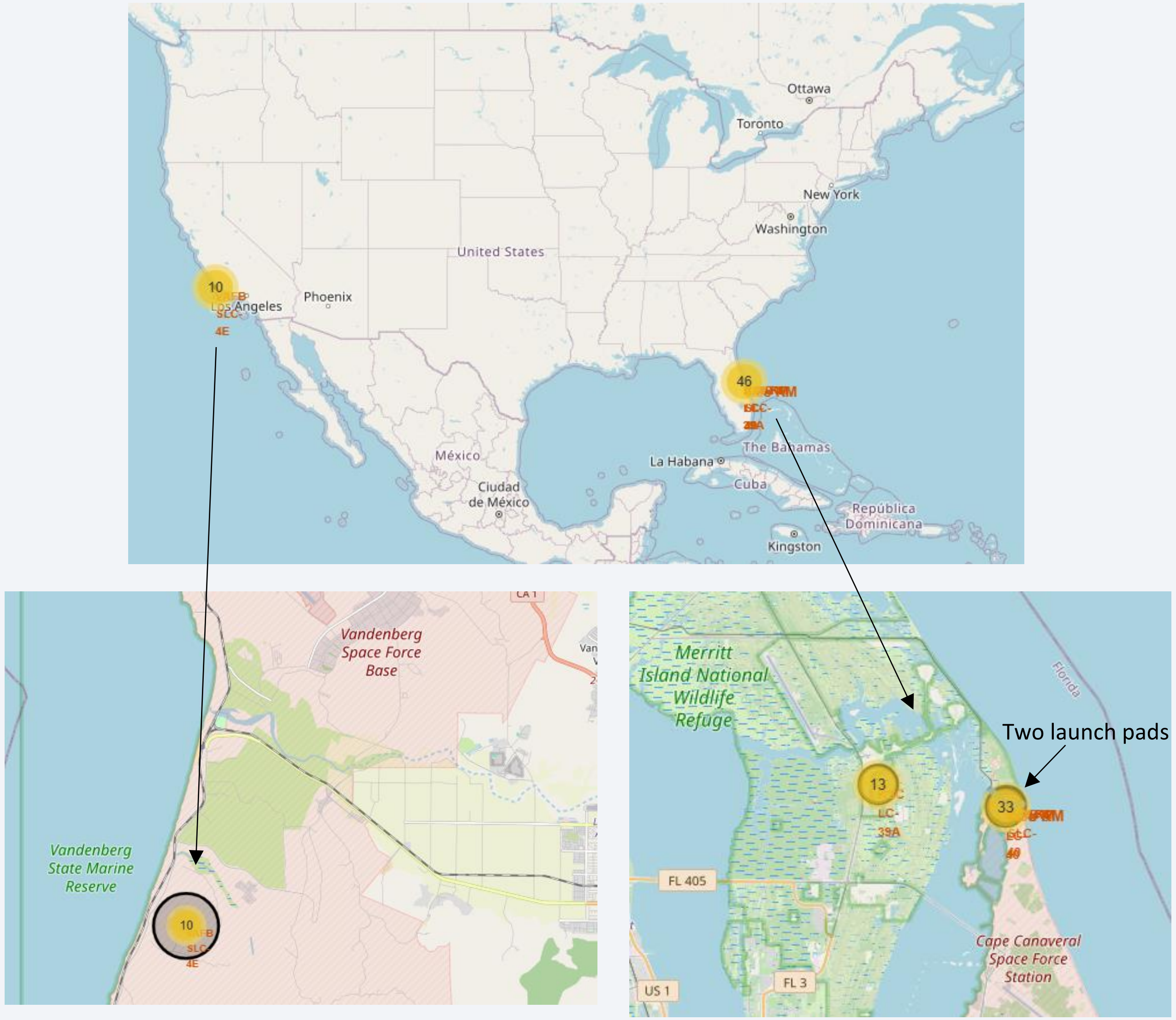
Insights from SQL queries

<table>
<tr><th>Data exploration</th><th>SQL Query</th><th>Query Result</th></tr>
<tr>
<td>Number of successful and failed mission outcomes. (Note: the initial code returned "Success" twice so the TRIM and LOWER statements are used to normalize the Mission_Outcome column and remove any whitespace and ensure case-insensitive grouping).</td>
<td>%%sql
SELECT TRIM(LOWER(Mission_Outcome)) AS Mission_Outcome, COUNT(*) AS Total FROM SPACEXTABLE
GROUP BY TRIM(LOWER(Mission_Outcome))</td>
<td>
<table>
<tr><th>Mission_Outcome</th><th>Total</th></tr>
<tr><td>failure (in flight)</td><td>1</td></tr>
<tr><td>success</td><td>99</td></tr>
<tr><td>success (payload status unclear)</td><td>1</td></tr>
</table></td>
</tr>
<tr>
<td>Names of the booster that has carried the maximum payload mass.</td>
<td>%sql select Booster_Version from SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (select Max(PAYLOAD_MASS__KG_) from SPACEXTABLE)</td>
<td>
<table>
<tr><th>Booster_Version</th></tr>
<tr><td>F9 B5 B1048.4</td></tr>
<tr><td>F9 B5 B1049.4</td></tr>
<tr><td>F9 B5 B1051.3</td></tr>
<tr><td>F9 B5 B1056.4</td></tr>
<tr><td>F9 B5 B1048.5</td></tr>
<tr><td>F9 B5 B1051.4</td></tr>
<tr><td>F9 B5 B1049.5</td></tr>
<tr><td>F9 B5 B1060.2</td></tr>
<tr><td>F9 B5 B1058.3</td></tr>
<tr><td>F9 B5 B1051.6</td></tr>
<tr><td>F9 B5 B1060.3</td></tr>
<tr><td>F9 B5 B1049.7</td></tr>
</table></td>
</tr>
<tr>
<td>Dates, booster version, and launch site in 2015 where there was an unsuccessful landing on the drone ship.</td>
<td>%%sql select Date, Booster_Version, Launch_Site, Landing_Outcome from SPACEXTABLE
WHERE Landing_Outcome = 'Failure (drone ship)' AND substr(Date,0,5)='2015’</td>
<td>
<table>
<tr><th>Date</th><th>Booster_Version</th><th>Launch_Site</th><th>Landing_Outcome</th></tr>
<tr><td>2015-01-10</td><td>F9 v1.1 B1012</td><td>CCAFS LC-40</td><td>Failure (drone ship)</td></tr>
<tr><td>2015-04-14</td><td>F9 v1.1 B1015</td><td>CCAFS LC-40</td><td>Failure (drone ship)</td></tr>
</table></td>
</tr>
<tr>
<td>The various landing outcomes between June 4, 2010, and March 20, 2017 (ranked in descending order).</td>
<td>%%sql SELECT Date, Landing_Outcome, COUNT(*) AS count
FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY count DESC;</td>
<td>
<table>
<tr><th>Date</th><th>Landing_Outcome</th><th>count</th></tr>
<tr><td>2012-05-22</td><td>No attempt</td><td>10</td></tr>
<tr><td>2016-04-08</td><td>Success (drone ship)</td><td>5</td></tr>
<tr><td>2015-01-10</td><td>Failure (drone ship)</td><td>5</td></tr>
<tr><td>2015-12-22</td><td>Success (ground pad)</td><td>3</td></tr>
<tr><td>2014-04-18</td><td>Controlled (ocean)</td><td>3</td></tr>
<tr><td>2013-09-29</td><td>Uncontrolled (ocean)</td><td>2</td></tr>
<tr><td>2010-06-04</td><td>Failure (parachute)</td><td>2</td></tr>
<tr><td>2015-06-28</td><td>Precluded (drone ship)</td><td>1</td></tr>
</table></td>
</tr>
</table>

Section 3

Launch Sites Proximities Analysis

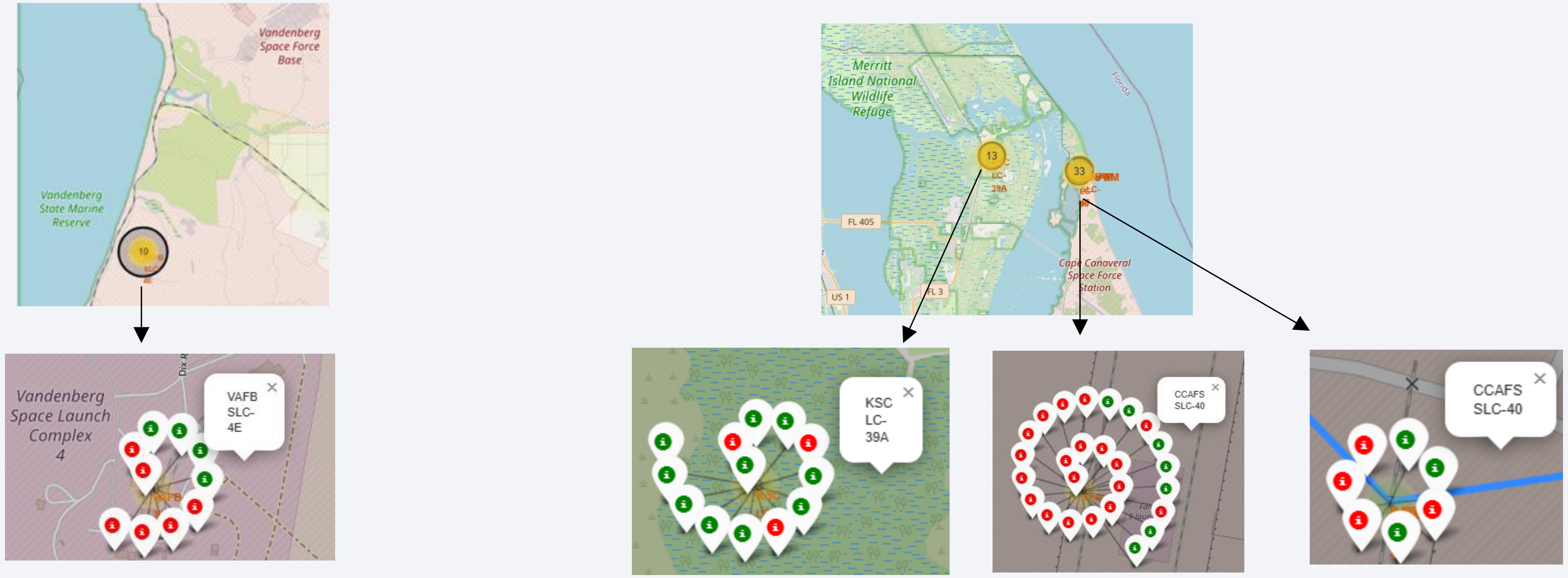
Falcon 9 launch site location



- The SpaceX dataset for the geo-mapping includes 56 flight records from four launch sites.
- The site locations are shown on screenshots of the map that was developed as part of the analysis (the numbers in the circles indicate the number of launches from the site).
- Three of the sites are located in Florida and include the Cape Canaveral Space Force Station with two launch pads (CCAFS LC-40 and CCAFS SLC-40) and the Kennedy Space Station (KSC LC-39A).
- The other launch site is the Vandenberg Space Force Base (VAFB SLC-4E) in California.
- The interactive map can be viewed at this link: [Launch site_map.html]

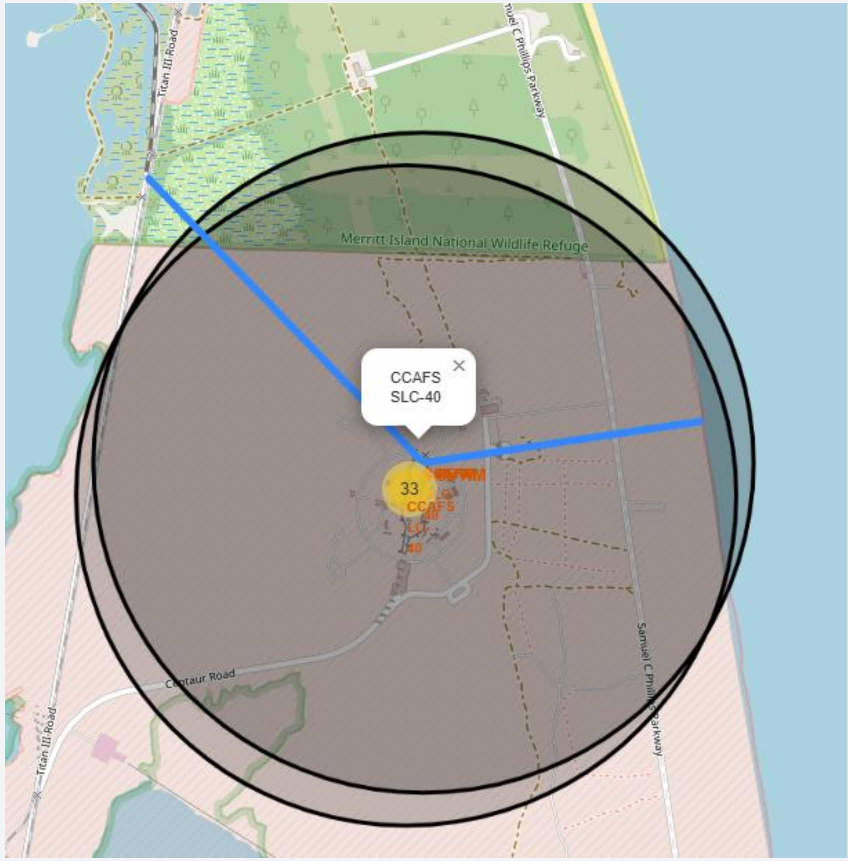
Note:

Launch site success rate

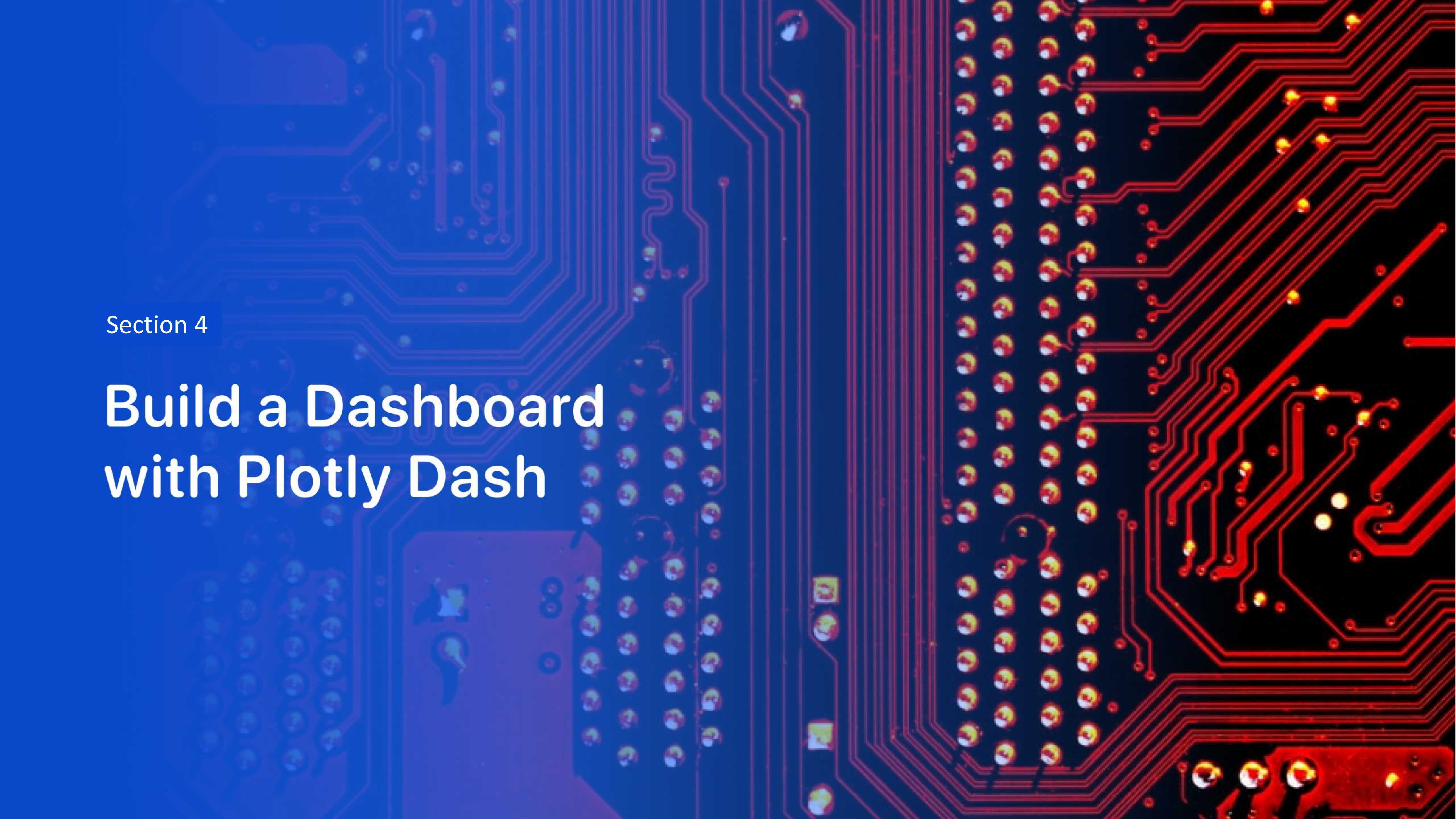

Vandenberg Space Force Base
Vandenberg State Marine Reserve
10
Merritt Island National Wildlife Refuge
13
LC-39A
33
FL 405
US 1
FL 3
Cape Canaveral Space Force Station
Vandenberg Space Launch Complex 4
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
CCAFS SLC-40

- As part of the launch site geo analysis, the landing success rate for each site was plotted on the map with markers (green is a successful landing and red is an unsuccessful landing).
- Overall, it is interesting to note that there are fewer successful landings than unsuccessful ones (see also dashboard slides below). This is largely due to the high failure rate at one of the Cape Canaveral launch pads.
- At the site level, the Kennedy Space Station (LC-39A) site has the highest success rate with 77% (10 out of 13 flights) of the Falcon flights landing successfully.
- As mentioned above, one of the Cape Canaveral launch pads (LC-40) has the lowest success rate with just 27% of the flights landing successfully.

Launch site proximities


CCAFS
SLC-40
33
Merritt Island National Wildlife Refuge
Titan III Road
Samuel C Phillips Parkway
Centaur Road

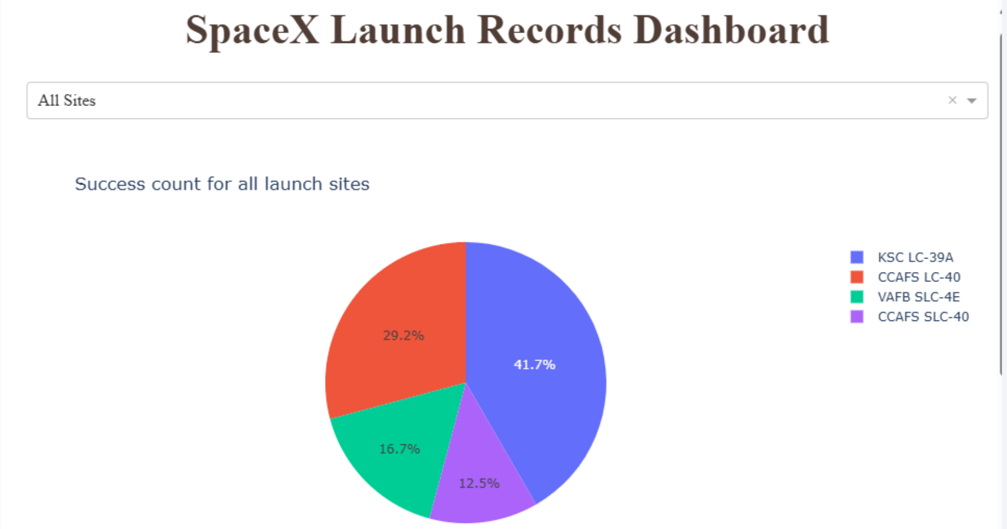
- As part of the launch site geo exploration and analysis, distances from the launch sites to areas of interest were calculated and displayed on the map.
- As an example, the Cape Canaveral launch pad SLC-40 is about 1 kilometer from the nearest railway and coastline.



Section 4

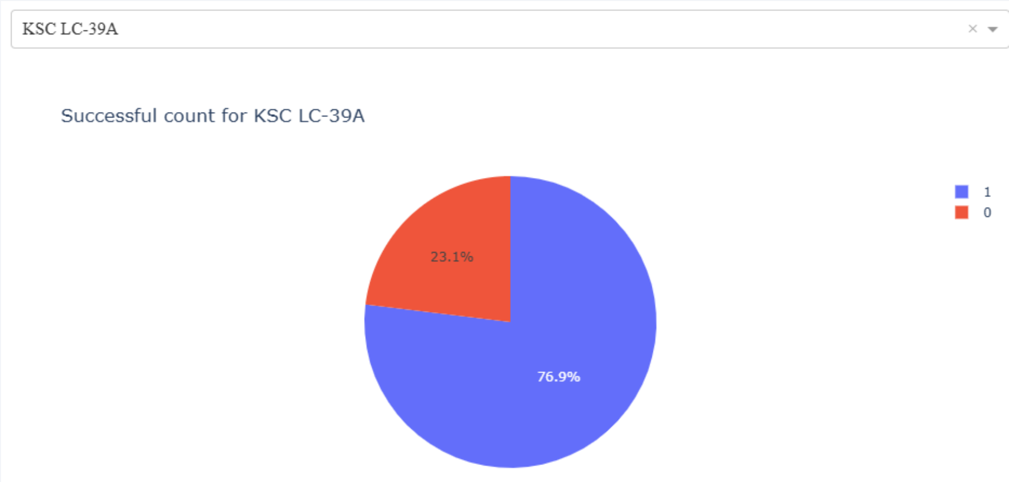
Build a Dashboard with Plotly Dash

SpaceX Dashboard-landing success counts and rates



SpaceX Launch Records Dashboard

All Sites

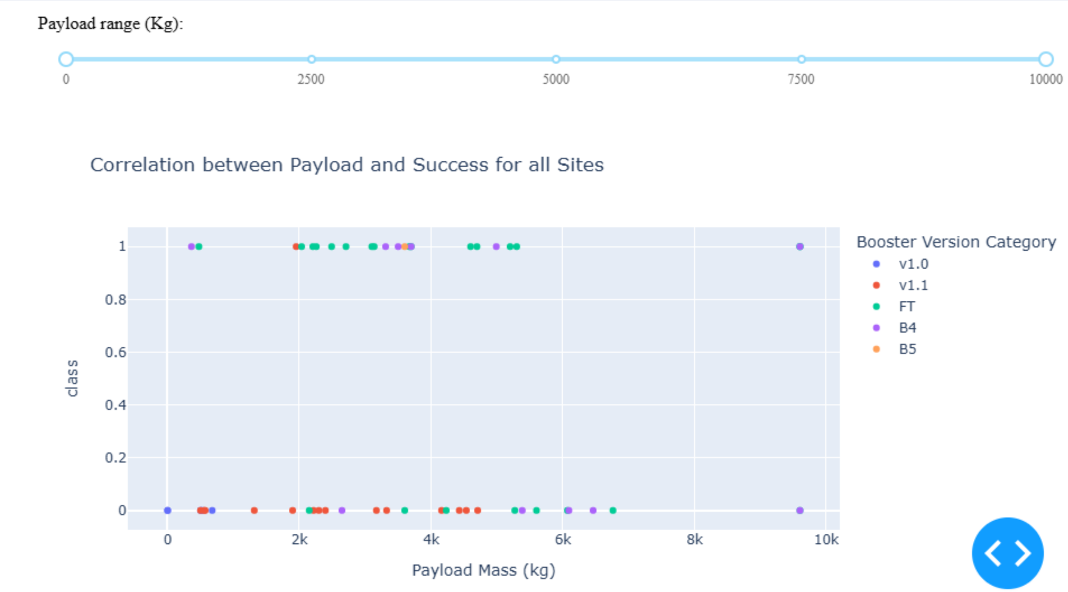


KSC LC-39A

- As mentioned previously a dashboard was developed for additional insights using the same dataset as above (for mapping).
- The top dashboard chart on the left is showing each launch site's percentage of the total (24) successful landings for all sites during the period.
- For example, of the 24 successful landings, the Kennedy Space Station site (KSC LC-39A) has the largest percentage of landings (10/24) while one of the Cape Canaveral sites (CCAFS SLC-40) has the lowest (3/24).
- Filtering the top chart by site shows the success rate – the ratio of successful landings to unsuccessful ones – for that site.
- As an example, the lower pie chart is filtered on the Kennedy Space Station site (KSC LC-39A), which as noted earlier has the highest success rate of 77%.

The dashboard can be viewed in a Colab notebook at the link below (after opening the link, click `Runtime` > `Run all` to start the dashboard and scroll down).
https://colab.research.google.com/drive/1Z-3E-rzsaXiOeCD-bTwITxZPz_qsaTia?usp=sharing

SpaceX Dashboard – success and payload by booster type

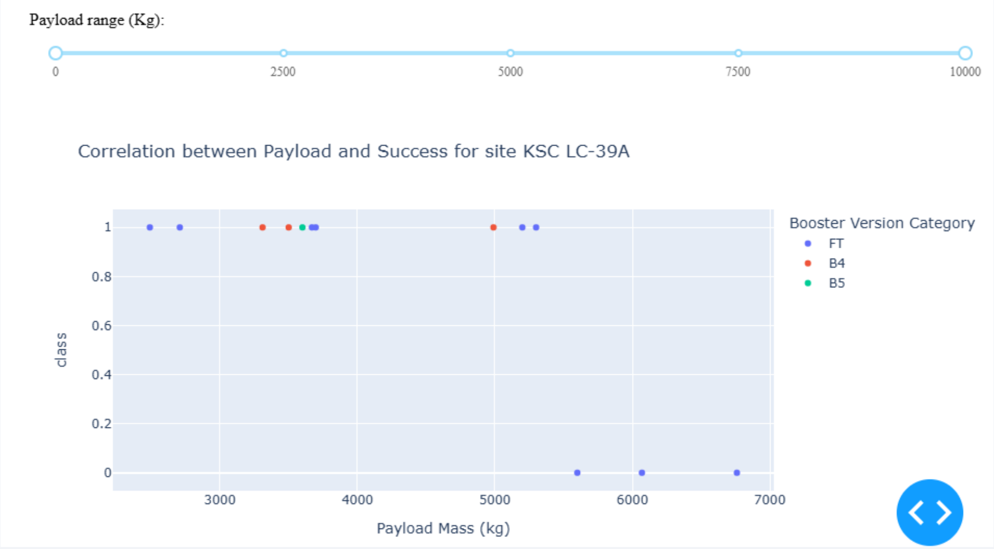


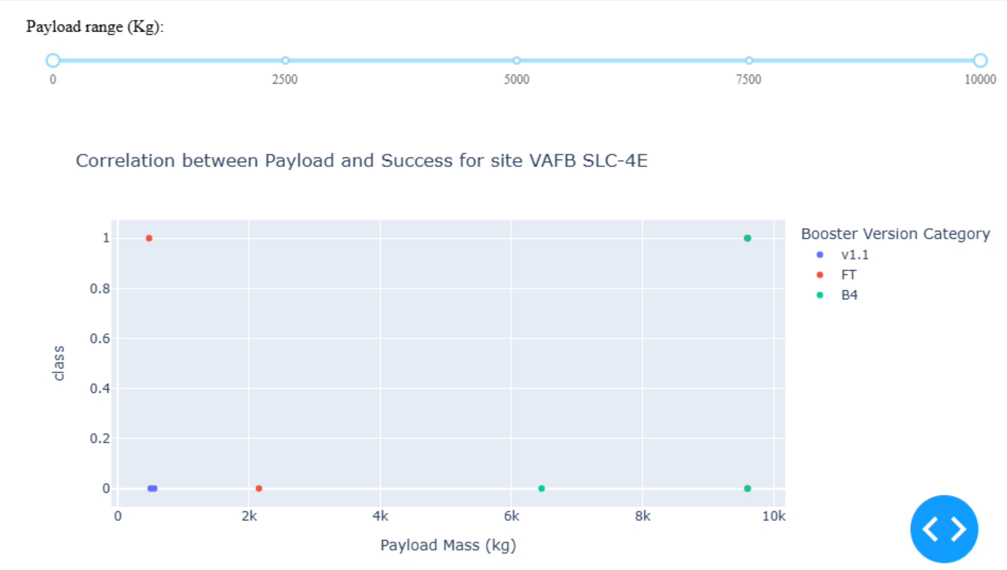
The dashboard also shows the relationship between flight payload and landing success rates for different Falcon 9 booster categories. The data can be filtered by launch site using the same filter as above for the pie chart (previous slide) and a *Payload range (Kg)* slider allows the user to filter the data through a range of different payloads.

- Overall, the Falcon FT booster has the most flights during the period and one of the highest successful landing rates at 67% (other than booster B5 which is 100% but with just one flight).
- On the flip side, the booster v1.0 has significantly fewer flights and no successful landings. Filtering the data by launch site shows that the v1.0 flights were made from the LC-40 site.
- Booster v1.1 also has a low landing success rate of 7.0%.
- Statistical modeling indicates that the booster version is the strongest predictor of a Falcon 9 landing outcome.
- Turning to payloads, the largest number of successful landings have had payloads ranging between 2,500 -5,000kg.
- There have been relatively few flights and successful landings with heavier payloads over 5,000kg.

The dashboard can be viewed in a Colab notebook at the link below (after opening the link, click `Runtime` > `Run all` to start the dashboard and scroll down).
https://colab.research.google.com/drive/1Z-3E-rzsaXiOeCD-bTwITxZPz_qsaTia?usp=sharing

SpaceX Dashboard – success and payload by booster type





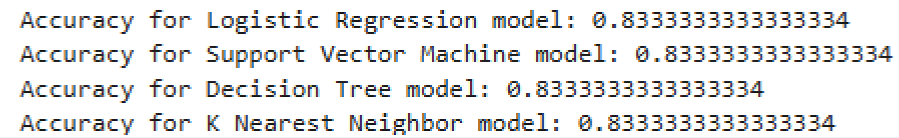
- Filtering the data by launch site shows interesting patterns and relationships not evident at the overall/ aggregate level.
- For example, the top chart on the left shows the data for the Kennedy Space Centre (LC-39A) – the site with the highest landing success rate overall as noted previously.
- Most of the site's successful landings were with relatively lighter payloads, while flights with heavier payloads (5,500kg. +) landed unsuccessfully.
- Switching to the Vandenberg Space Force Base site (SLC-4E), the bottom chart shows that this launch site has had successful landings with the heavier payloads (note: although the chart is showing one successful landing for a payload of 9,600kg. there are actually three successful landings with this payload-the chart dots are stacked on top of each other).
- Interestingly, of Vandenberg's four successful landings, three-quarters of them have been with heavy payloads.

The dashboard can be viewed in a Colab notebook at the link below.
https://colab.research.google.com/drive/1Z-3E-rzsaXiOeCD-bTwITxZPz_qsaTia?usp=sharing

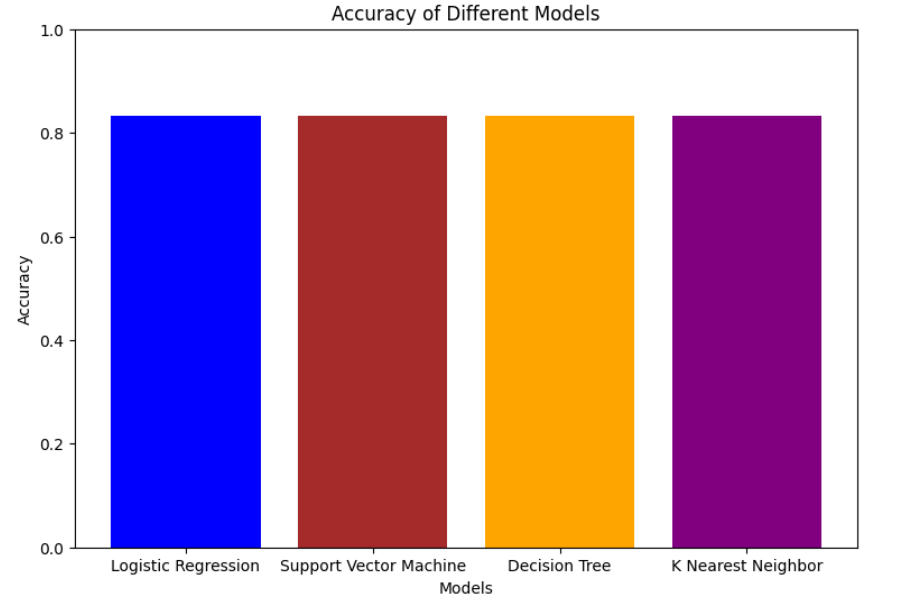
Section 5

Predictive Analysis (Classification)

Classification Model Accuracy



```
Accuracy for Logistic Regression model: 0.8333333333333334
Accuracy for Support Vector Machine model: 0.8333333333333334
Accuracy for Decision Tree model: 0.8333333333333334
Accuracy for K Nearest Neighbor model: 0.8333333333333334
```



- As mentioned earlier four classification models were developed for predicting Falcon 9 landing outcomes: Logistic Regression, K-Nearest Neighbor, Decision Tree and Support Vector Machine.

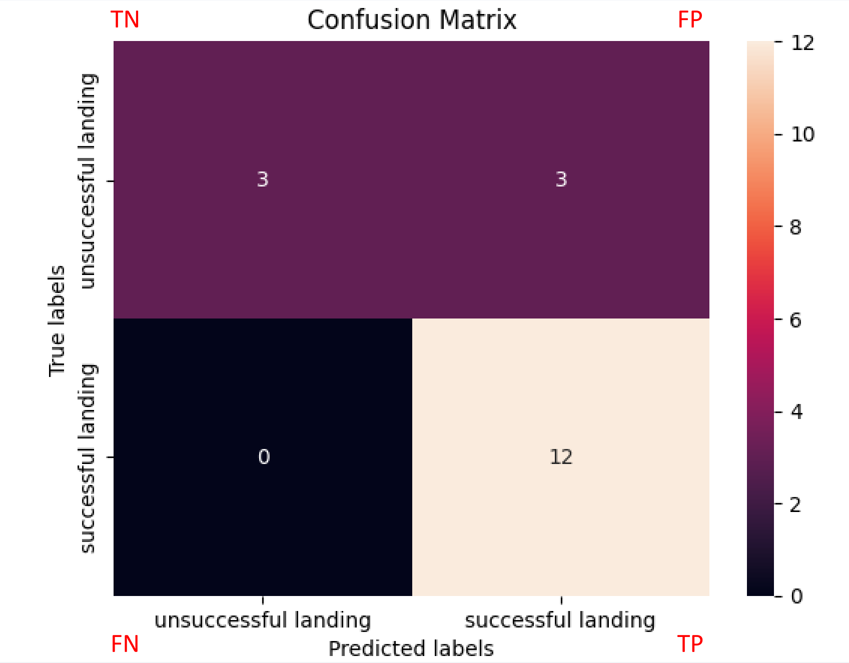
- One of the metrics used to evaluate the predictive performance of the models is python's *sklearn accuracy score*.

- The accuracy score measures the percentage of correct predictions, or in this case the landing outcomes that were correctly classified by the model algorithms.

- The table and chart show that the four models have the same accuracy score of 0.83.

* The models correctly predicted the Falcon 9 landing outcome for 83% of the flights.

Confusion Matrix



- The sklearn's *Confusion Matrix* is another measure that provides additional information about the model's accuracy in predicting a Falcon 9 landing outcome.[4]
- It can be noted from the matrix that out of the test sample data of 18 flights, 12 landings were actually successful and 6 were unsuccessful ('True labels').
- The models predicted 12 successful landings and 3 unsuccessful ones ('Predicted labels').

* The models' main prediction weakness is generating false positives (FP), i.e., predicting a successful landing when there isn't one.

[4] The rows in the matrix represent the *actual* landing outcomes (training data) and the columns represent the *predicted* landing outcomes (test data). Accuracy = TP + TN / total = (15/18* 100 = 83.3%)

Model Choice

- Choosing the best predictive model to use when they have the same performance metrics such as accuracy score and confusion matrix, largely depends on model interpretability and complexity - it is important to understand what the model is doing and be able to interpret and apply the results.
- Other factors when deciding on which model to use are computing resources and computational cost, the size of the dataset and scalability, and general robustness to noisy data, model over/underfitting, and generalization.
- In view of the above and for predicting SpaceX rocket landing outcomes, the logistic regression model is a suitable choice that offers reasonable interpretability and a clear relationship between landing outcome and the features/predictors.
- For example, with some transformation, the model's regression coefficients can be interpreted as an odds ratio to give the odds of a successful landing based on various predictors, and whether the odds increase or decrease with changes in the predictors.[5]
- A preliminary SpaceX logistic regression model and coefficients is shown on the next slide – the model is useful for predicting the probability of successful landings for future Falcon 9 flights based on these flight characteristics and predictors:

[5] The logistic regression coefficient (β) is the estimated change in the log odds of the response variable (landing outcome) associated with a one-unit change in the feature (predictor) variable. Exponentiating the coefficient (e^{β}) transforms the results to an odds ratio for easier interpretation.

SpaceX Logistic Regression Model

Feature Category	Regression Coefficient (β)	Odds Ratio (e^{β})
BoosterVersion	1.341	3.825
Legs	0.243	1.275
GridFins	0.218	1.244
Orbit	0.217	1.242
LandingPad	0.155	1.168
ReusedCount	0.078	1.081
LaunchSite	0.055	1.056
Block	0.053	1.055
FlightNumber	0.048	1.049
Reused	0.040	1.040
PayloadMass	0.024	1.024
Flights	0.001	1.001

B_0 intercept = 0.7904

- The *Booster Version* is the strongest predictor of a falcon 9 landing outcome. For example, a one-unit difference in the booster version almost triples the odds of a successful landing (i.e., a 282.5% increase).

- Other predictors such as *Legs*, *Grid Fins*, *Orbits*, and *landing pads* also have a significant affect on landing outcomes (16.8% - 27.5%).

- Payload Mass (kg.) and Flights have a negligible impact on the success of a landing.

- The probability of a successful landing for a future Falcon 9 flight can be estimated by entering the values for the flight into the regression model below (for example, a flight using a particular booster, launching from one of the SpaceX launch sites, and going to a certain orbit).

- Further model development and refinement should include confirming the model assumptions (e.g. independence, multicollinearity) and examining the statistical significance (p-value) and standard errors of the coefficients.

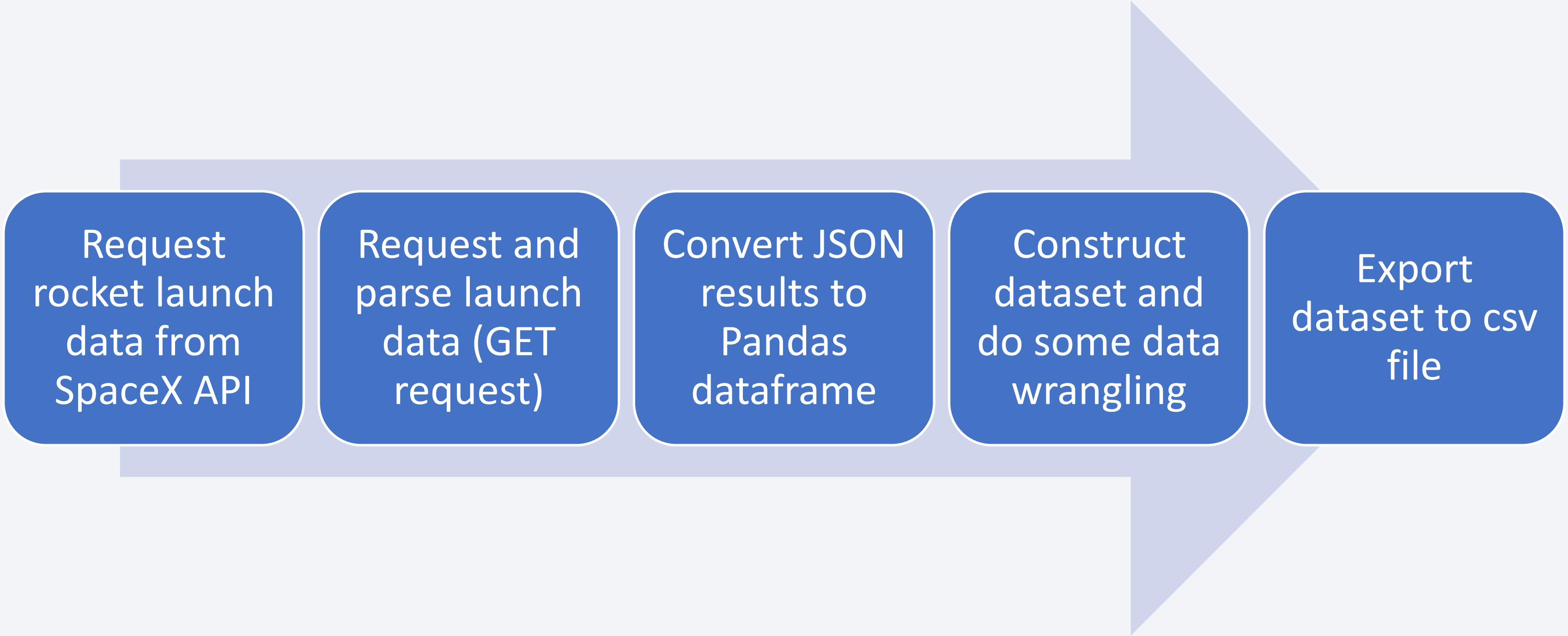
Model

$$\text{Log}\left[\frac{P(Y=1)}{1-P(Y=1)}\right] = 0.7904 + 1.341 \times \mathit{BoosterVersion} + 0.243 \times \mathit{Legs} + 0.218 \times \mathit{GridFins} + 0.217 \times \mathit{Orbit} + 0.155 \times \mathit{LandingPad} + 0.078 \times \mathit{ReusedCount} + 0.055 \times \mathit{LaunchSite} + 0.053 \times \mathit{Block} + 0.048 \times \mathit{FlightNumber} + 0.040 \times \mathit{Reused} + 0.024 \times \mathit{PayloadMass} + 0.001 \times \mathit{Flights}$$

Conclusion

- The overall landing success rate for the Falcon 9 rocket during the period is 67% although this varies based on a number of factors/predictors.
- The Kennedy Space Station (LC-39A) site has the highest landing success rate with 77% (10 out of 13 flights) of the Falcon flights landing successfully. Launch sites have a relatively small influence on rocket landing outcomes in the machine learning models.
- During the 10-year period between 2010 - 2020 the Falcon 9 average landing success rate has been increasing.
- Heavier payloads up to a certain weight may increase the likelihood of successful landings. The largest number of successful landings have payloads ranging between 2,500 -5,000kg. Overall however, statistical modeling indicates that payload has a relatively small effect on landing outcomes.
- Regression modeling indicates that SpaceX orbits have a notable impact on the likelihood of a successful landing. Excluding orbits with just one flight, flights to the sun-synchronous orbit (SSO) have a perfect landing success rate while the Very Low Earth Orbits (VLEO) also have relatively high landing success rates (85.5%).
- The Falcon FT booster has the most flights during the period and the highest successful landing rate at 67% (excluding boosters with just one flight). Meanwhile, the v1.0 booster has no successful landings. Statistical modeling shows that the booster version is the strongest predictor of Falcon 9 landing outcomes.
- Four machine learning models were developed during the study: Logistic Regression, K-Nearest Neighbor, Decision Tree, and Support Vector Machine. In terms of model performance, they have the same accuracy score and confusion matrix.
- The models correctly predicted the Falcon 9 landing outcome for 83% of the flights during the study period.
- The models main prediction weakness is generating false positives (FP), i.e., predicting a successful landing when there isn’t one.
- In choosing one of the models, the logistic regression model is a suitable choice that offers reasonable interpretability and a clear relationship between Falcon 9 landing outcomes and the flight features/predictors.
- A preliminary logistic regression model has been developed which can be used to predict the probability of successful landings for future Falcon 9 flights. Model refinement and improvement can include training the model on a larger data set, reviewing the regression coefficients and feature selection/engineering, and bringing in domain knowledge and expertise.

Appendix 1. REST API Process



Request rocket launch data from SpaceX API

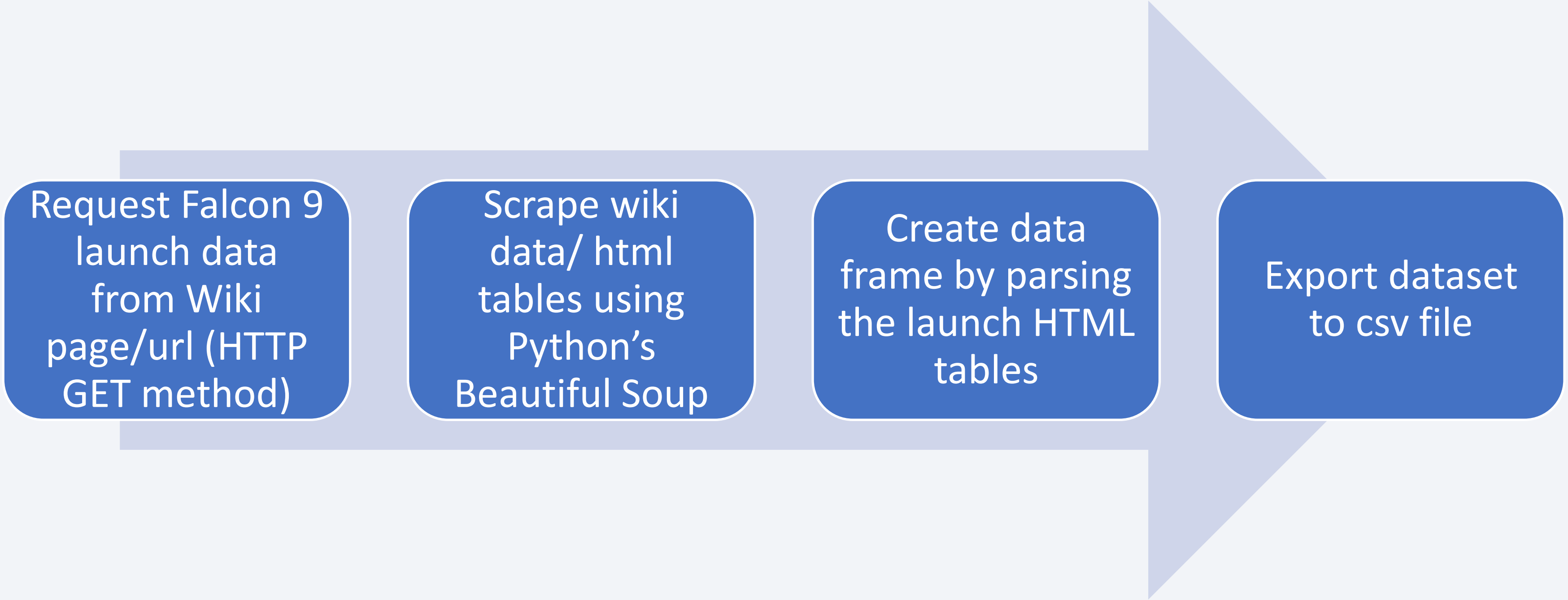
Request and parse launch data (GET request)

Convert JSON results to Pandas dataframe

Construct dataset and do some data wrangling

Export dataset to csv file

Appendix 2. Web Scraping Process



Request Falcon 9 launch data from Wiki page/url (HTTP GET method)

Scrape wiki data/ html tables using Python’s Beautiful Soup

Create data frame by parsing the launch HTML tables

Export dataset to csv file

Appendix 3. Rocket Booster Landing Outcomes

Booster Landing	Outcome	# Landings
False ASDS	Unsuccessful landing on Autonomous Spaceport Drone Ship (ASDS)	6
False Ocean	Unsuccessful ocean test (uncontrolled landing, no recovery)	2
False RTLS	Unsuccessful Return to Launch Landing Site (RTLS)	1
None ASDS	Precluded, drone ship (launch-related problems)	2
None None	No attempt	19
True ASDS	Successful landing on Autonomous Spaceport Drone Ship (ASDS)	41
True Ocean	Successful ocean test (controlled landing, no recovery)	5
True RTLS	Successful Return to Launch Landing Site (RTLS)	14

Note: For statistical modeling purposes, the first five landing categories in the table are considered *unsuccessful landings* and assigned a value of 0 in a derived variable *Class*, while the remaining three categories are considered *successful landings* and assigned a value of 1 in *Class*.

Appendix 4. SpaceX Orbits

Orbit	Description
ES-L1	At the Lagrange points the gravitational forces of the two large bodies cancel out in such a way that a small object placed in orbit there is in equilibrium relative to the center of mass of the large bodies. L1 is one such point between the sun and the earth [5] .
GEO	This is a circular geosynchronous equatorial orbit 35,786 kilometres (22,236 miles) above Earth's equator and following the direction of Earth's rotation [10].
GTO	A geosynchronous orbit is a high Earth orbit that allows satellites to match Earth's rotation. Located at 22,236 miles (35,786 kilometers) above Earth's equator, this position is a valuable spot for monitoring weather, communications and surveillance. Because the satellite orbits at the same speed that the Earth is turning, the satellite seems to stay in place over a single longitude, though it may drift north to south," NASA wrote on its Earth Observatory website [3] .
HEO	A highly elliptical orbit, is an elliptic orbit with high eccentricity, usually referring to one around Earth [6].
ISS	A modular space station (habitable artificial satellite) in low Earth orbit. It is a multinational collaborative project between five participating space agencies: NASA (United States), Roscosmos (Russia), JAXA (Japan), ESA (Europe), and CSA (Canada) [7].
LEO	Low Earth orbit (LEO) is an Earth-centred orbit with an altitude of 2,000 km (1,200 mi) or less (approximately one-third of the radius of Earth),[1] or with at least 11.25 periods per day (an orbital period of 128 minutes or less) and an eccentricity less than 0.25.[2] Most of the manmade objects in outer space are in LEO [1].
MEO	Geocentric orbits ranging in altitude from 2,000 km (1,200 mi) to just below geosynchronous orbit at 35,786 kilometers (22,236 mi). Also known as an intermediate circular orbit. These are "most commonly at 20,200 kilometers (12,600 mi), or 20,650 kilometers (12,830 mi), with an orbital period of 12 hours [8].
PO	Polar orbit is one type of satellites in which a satellite passes above or nearly above both poles of the body being orbited (usually a planet such as the Earth [11].
SO	
SSO	A Sun-synchronous orbit, also called a heliosynchronous orbit, is a nearly polar orbit around a planet, in which the satellite passes over any given point of the planet's surface at the same local mean solar time [4] .
VLEO	Very Low Earth Orbits (VLEO) can be defined as the orbits with a mean altitude below 450 km. Operating in these orbits can provide a number of benefits to Earth observation spacecraft as the spacecraft operates closer to the observation[2].

Source: IBM Data Science Capstone project; Wikipedia.

Appendix 5 - Model Parameters

Model	Parameters (Python dictionary)	Parameter Description

Logistic regression	parameters ={'C':[0.01,0.1,1],
	'penalty':['l2'],
	'solver':['lbfgs']}
		Control the type and strength of regularization (overfitting) and specify which logistic regression algorithm to use.

K-Nearest Neighbor	parameters = {'n_neighbors': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],
	'algorithm': ['auto', 'ball_tree', 'kd_tree', 'brute'],
	'p': [1,2]}
		Customize the KNN model such as the number of neighbors, algorithm used, and distance metric.

Decision Trees	parameters = {'criterion': ['gini', 'entropy'], 'splitter': ['best', 'random'], 'max_depth': [2*n for n in range(1,10)], 'max_features': ['log2', 'sqrt’],# auto not included due to partitioning errors 'min_samples_leaf': [1, 2, 4], 'min_samples_split': [2, 5, 10]}	Determine how the decision tree will split the data based on the best feature, and at each node; the depth of the tree; the maximum number of features for splitting; and the minimum number of samples required for each leaf node and node split.

Support Vector Machine	parameters = {'kernel':('linear', 'rbf','poly','rbf', 'sigmoid'),
	'C': np.logspace(-3, 3, 5),
	'gamma':np.logspace(-3, 3, 5)}
		Specifying the type of SVM kernel, regularization strength (over/under-fitting), and decision boundary (kernel coefficient).

References

SpaceX website, Falcon 9: SpaceX - Falcon 9

Thank you!

Dave Plumstead
Email: dwplumst@gmail.com
LinkedIn: https://ca.linkedin.com/in/davidplumstead
GitHub: Your Repositories (github.com)
Tableau: https://public.tableau.com/app/profile/david.plumstead/vizzes