Winning Space Race with Data Science

Jeffrey Thomas
08 Dec 2022

Outline

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

Executive Summary

- The purpose of this study was to develop the capability to predict if the first stage of the SpaceX Falcon 9 rocket will successfully land.
- The landing of the first stage booster enables it to be reused for subsequent commercial space launches. The significant savings in costs provides SpaceX with a significant commercial advantage in attracting business.
- Open-source data (SpaceX REST API, Wikipedia) were sufficient to build a model for predicting whether a Falcon 9 booster will land successfully or not.
- A Decision-Tree Classifier was built, using input data from API and web-scraped data, with a 94% Accuracy Rate.

Introduction

The purpose of this study was to develop the capability to predict if the first stage of the SpaceX Falcon 9 rocket will successfully land. The landing of the used first stage booster enables the booster to be reused for subsequent commercial space launches. The significant savings in costs provides Spaces with a significant commercial advantage in attracting business.

Using publicly available data on SpaceX Falcon 9 launches, launch locations, orbits, payloads, and mission outcomes, a predictive model will be produced to determine the likelihood of the first stage successfully landing, and subsequently determine the cost of an upcoming launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data source 1: SpaceX Rest API
    - https://api.spacexdata.com/v4/launches/past
  - Data Source 2: Web Scraping
    - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- Perform data wrangling
  - Collected data was enriched by creating a Landing Outcome class variable based on performance data. This “class” variable would be predicted by the various machine learning models applied.

Methodology

Executive Summary

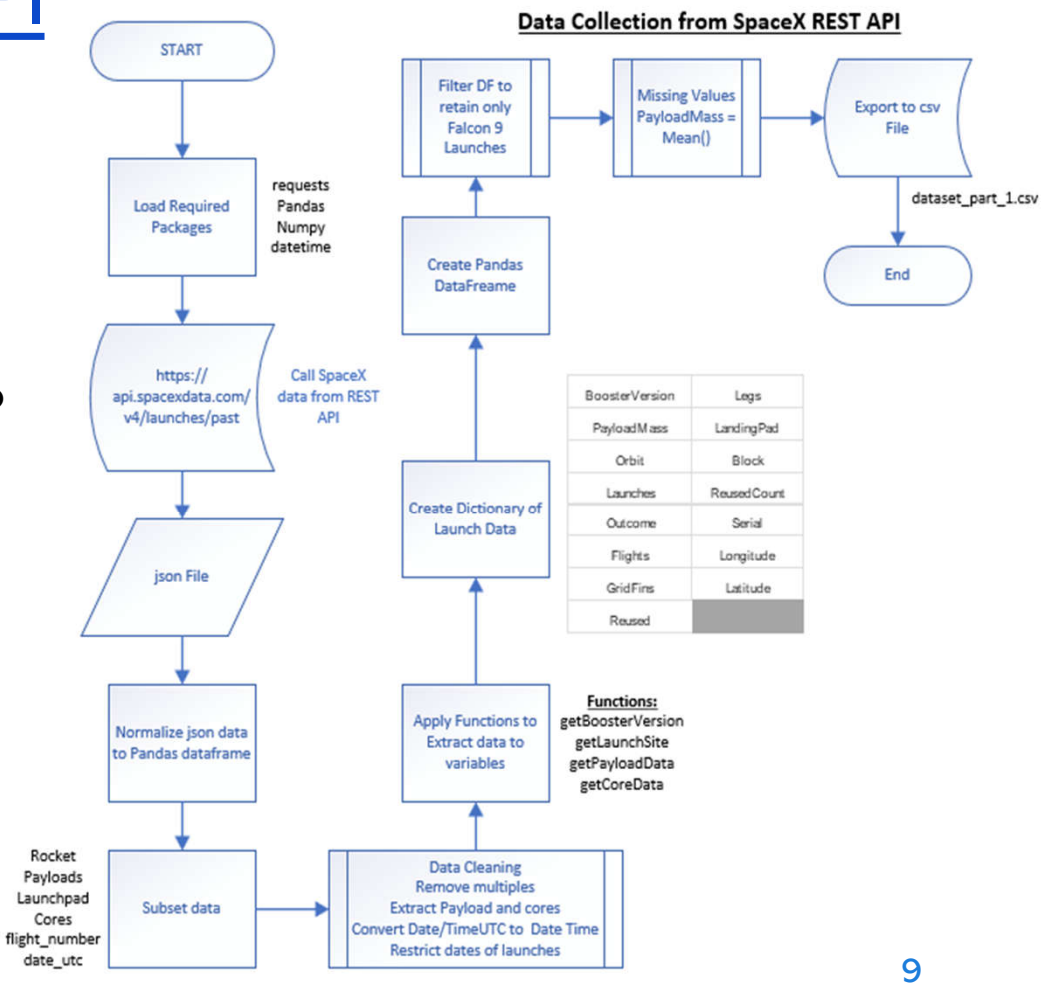
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Input data for the models were normalized, and split into Training and Testing datasets.
  - Models were optimized using a Cross-Validation Grid Search routine, to determine the best settings for each model's hyperparameters.

Data Collection

- Data for this study were collected from two primary sources:
  - Data source 1: SpaceX Rest API
    - https://api.spacexdata.com/v4/launches/past
  - Data Source 2:
    - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- Data collected from the SpaceX REST API were received in JSON format, and converted to a Pandas dataframe. The data were filtered to include only Falcon 9 data, missing values replaced, then exported to a csv file for further analysis.
- HTML data was web-scraped from the SpaceX Wikipedia page, and processed using the BeautifulSoup Python package, and converted to Pandas dataframe and exported to a csv file for further analysis.

Data Collection – SpaceX API

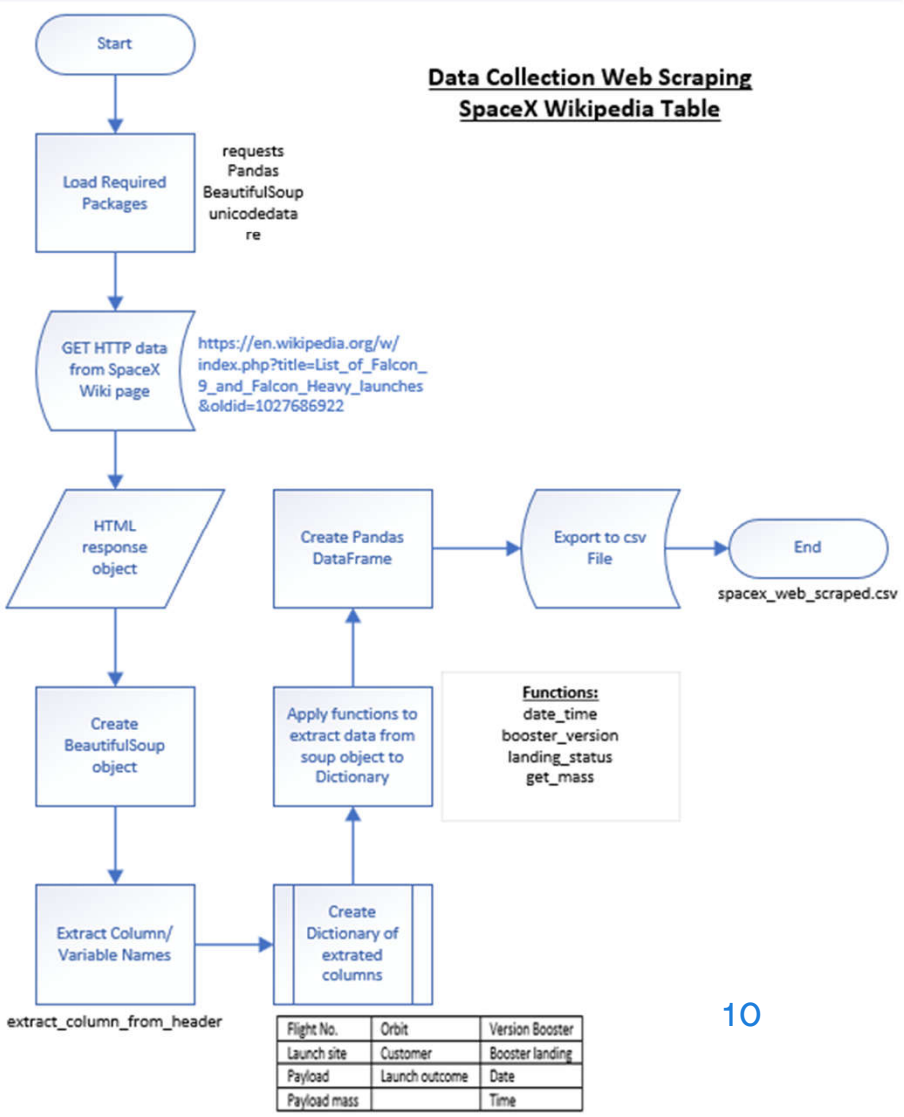
- Data is downloaded from the SpaceX REST API using the **requests.get()** call.
- JSON file is converted into a Pandas DataFrame object using the **json.normalize()** call.
- Data is subsetted using **data[].map** call.
- Functions (see flowchart) applied to extract variables to lists, which are combined into a dictionary.
- Convert dictionary into Pandas DataFrame.
- Filter data for only Falcon 9 data, replace missing PayloadMass data with mean of PayloadMass, using **replace()** call.
- Output data to csv file- dataset_part_1.csv
- Link to GitHub Jupyter Notebook:
  https://github.com/JTKE10901/Capstone-SpaceX-Project/blob/f8167ea5936bb43777cf0b00e37d3082632c6eab/SpaceX%20data%20with%20web%20scraping.ipynb



Data Collection – Web Scraping

- HTML Data is downloaded from the SpaceX Wikipedia table using the **requests.get().text** call.
- Use BeautifulSoup() call to create soup object.
- Apply functions (see flowchart) to extract data from soup object into Dictionary.
- Convert Dictionary into Pandas DataFrame.
- Export DataFrame to csv file- spacex_web_scraped.csv
- Link to GitHub Jupyter Notebook:

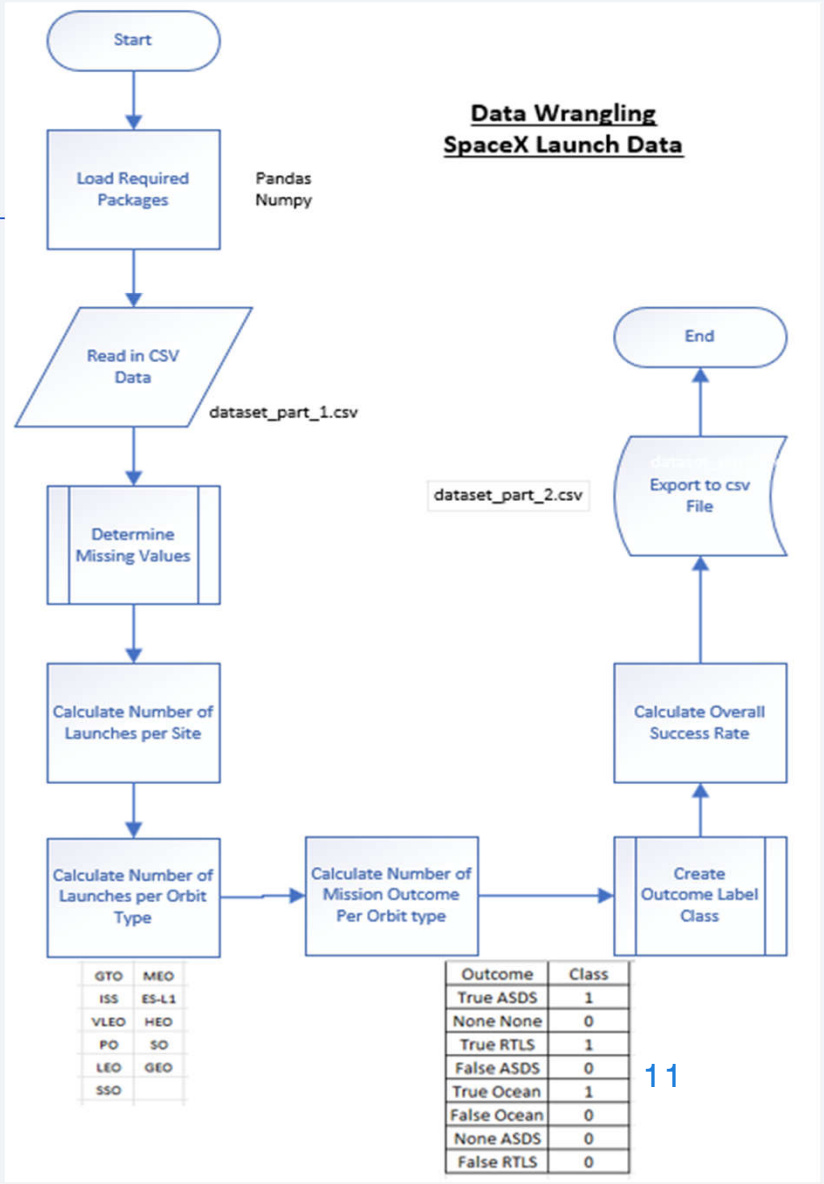
https://github.com/JTKE10901/Capstone-SpaceX-Project/blob/7c71d3c677115c2f9904c114c6d061a0da8ab503/SpaceX%20data%20with%20web%20scraping.ipynb



Data Wrangling

- Data were read in from csv file.
- The following values were calculated:
  - Percentage of missing values/variable.
  - Number of Launches per Site.
  - Number of Launches per Orbit Type.
  - Number of Mission Outcome Per Orbit type.
- Create Outcome Class categorical variable.
- Calculate Overall Success Rate
- Export data to csv file- dataset_part_2.csv
- Link to GitHub Jupyter Notebook:

https://github.com/JTKE10901/Capstone-SpaceX-Project/blob/4be8a5f8c39971fd9c27d494a694519c3685271d/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with SQL

- Names of the unique launch sites.
- Display 5 records where launch sites begin with the string 'KSC'.
- Display the total payload mass carried by boosters launched by NASA (CRS).
- Display average payload mass carried by booster version F9 v1.1.
- List the date where the first successful landing outcome in drone ship or ground pad.

Link to GitHub Juptyer Noterbook EDA with SQL

EDA with SQL (Cont'd)

- List the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000.
- List the total number of successful and failure mission outcomes.
- List the names of the booster_versions which have carried the maximum payload mass.
- List the records which will display the month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017.
- Rank the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

Link to GitHub Juptyer Noterbook EDA with SQL

EDA with Data Visualization

- The following plots were created to evaluate the SapceX Falcopn 9 booster performance:
  - Scatterplot of Payload Mass vs. Flight Number, to look at trends in Landing Outcomes over the course of SpaceX launches from 04 Jun 2010 through 05 Nov 2020, a total of 90 flights.
  - Scatterplot of Launch Site vs. Flight Number, to observe trends in Landing Outcomes over the course of SpaceX launches from 3 different launch sites.
  - Scatterplot of Payload vs. Launch Site, to determine if there are payload weight limits by launch site.

Link to GitHub Jupyter Notebook- EDA with Visualization

EDA with Data Visualization (Cont'd)

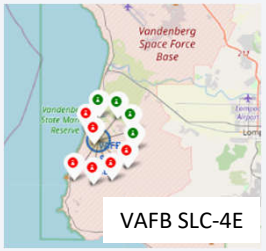
- Barchart to compare Success Rate of mission by Orbit Type.
- Scatterplot of Orbit type vs. FlightNumber, to see trends in orbit types over the course of time.
- Scatterplot of Payload vs. Orbit to see if Orbit Type has a relationship to Pay Load Mass and Mission Success.
- Line chart to evaluate the overall mission success rate over time (by Year).
- Conclude EDA with Features Engineering
  - Extract Features used in Models, convert to "one-hot encoded variables".
  - Export this data to csv file- "dataset_part_3.csv"

Link to GitHub Jupyter Notebook- EDA with Visualization

Build an Interactive Map with Folium

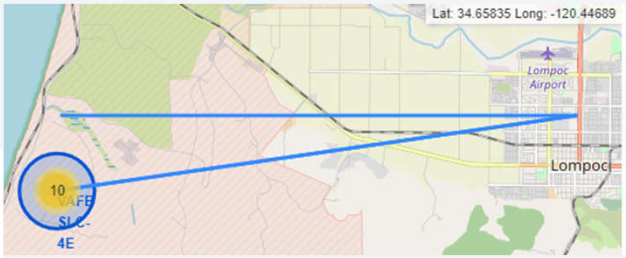
- Markers for each of the launch centers were placed on the map.
- Marker clusters were added in order to display the Success/Failure of each launch at each site.

Simple Marker at Johnson Space Control (JSC)



Marker Cluster at Vandenberg Space Launch Center

- Additionally, Latitude/Longitude display and distance lines were plotted to various landmarks to enable calculation of distances.



[Link to Juptyer Notebook with Folium Interactive Maps]- Only a portion of notebook displays. SEE Link to IBM Cloud version.

[Link to IBM Cloud Version]- This one should work.

Build a Dashboard with Plotly Dash

- A Dashboard was built, with a dropdown list of the available Launch Sites, a Pie chart showing % of successful launches by Launch Site or % Success vs % Failed launches for a specific Launch Site.
- A slider was added to enable the user to select ranges of payload mass, in addition to the Launch Site dropdown, to enable evaluation of performance under specific conditions.
- A scatterplot showing the correlation between payload and launch success.

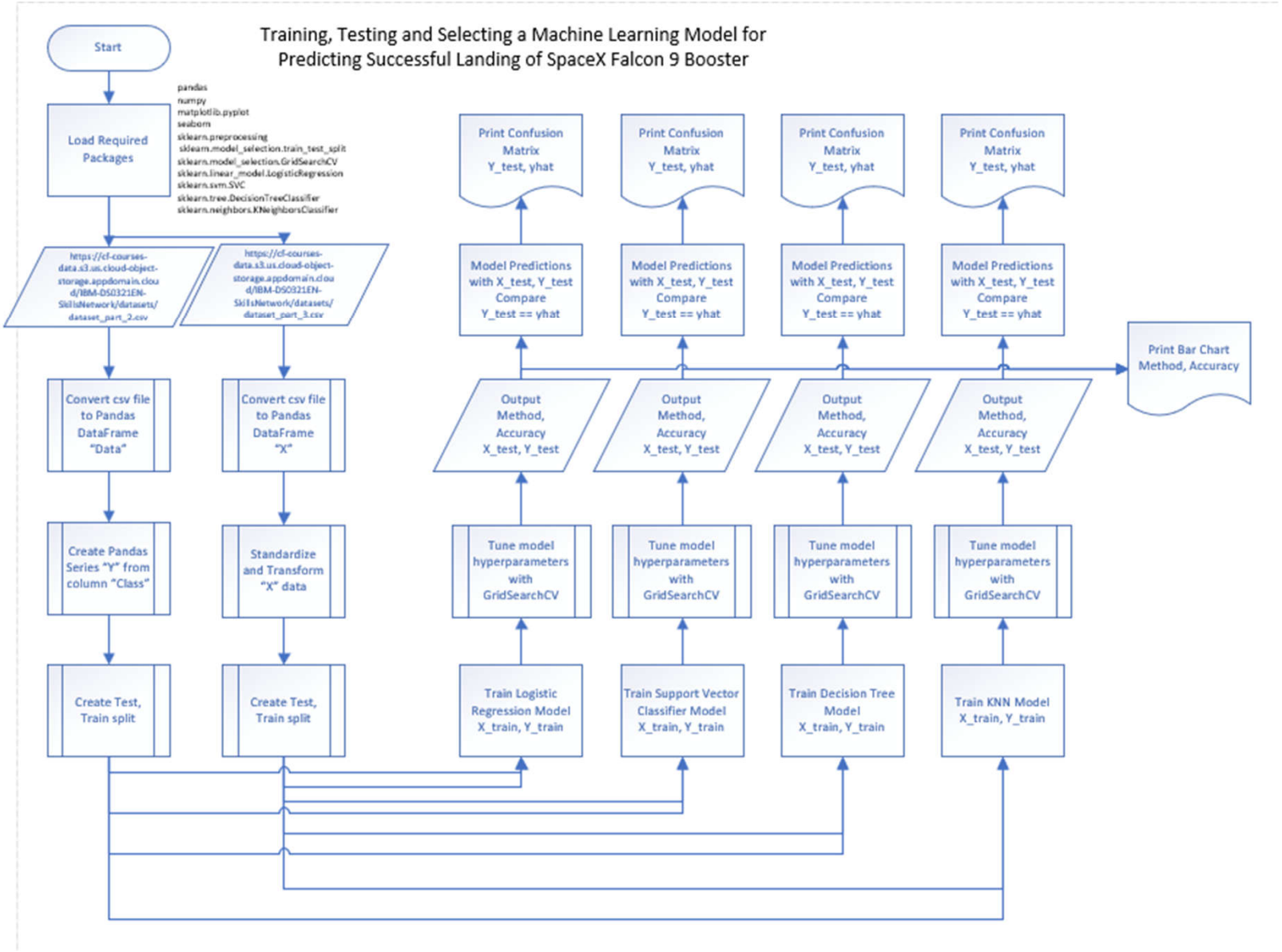
Link to GitHub SpaceX Dash App

Predictive Analysis (Classification)

- Data (dataset_part_2.csv and dataset_part_3.csv) were read into the program.
- Each of the csv files was converted to a Pandas DataFrame.
- Create Pandas series “Y” from dataset_part_2 (Class)
- Standardize and Transform DataFrame “X”.
- Create Train/Test splits from the data “X” and “Y”(80% Train:20% Test).
- Train 4 models with Training data
  - Logistic regression, Support vector machine, Decision tree classifier, K nearest neighbors
  - Use CVGridSearch to find optimized hyperparameters for each model.
- Fit “X_Test” data to optimized models to predict “yhat”, compare to Y_test.
- Calculate Accuracy for each model.

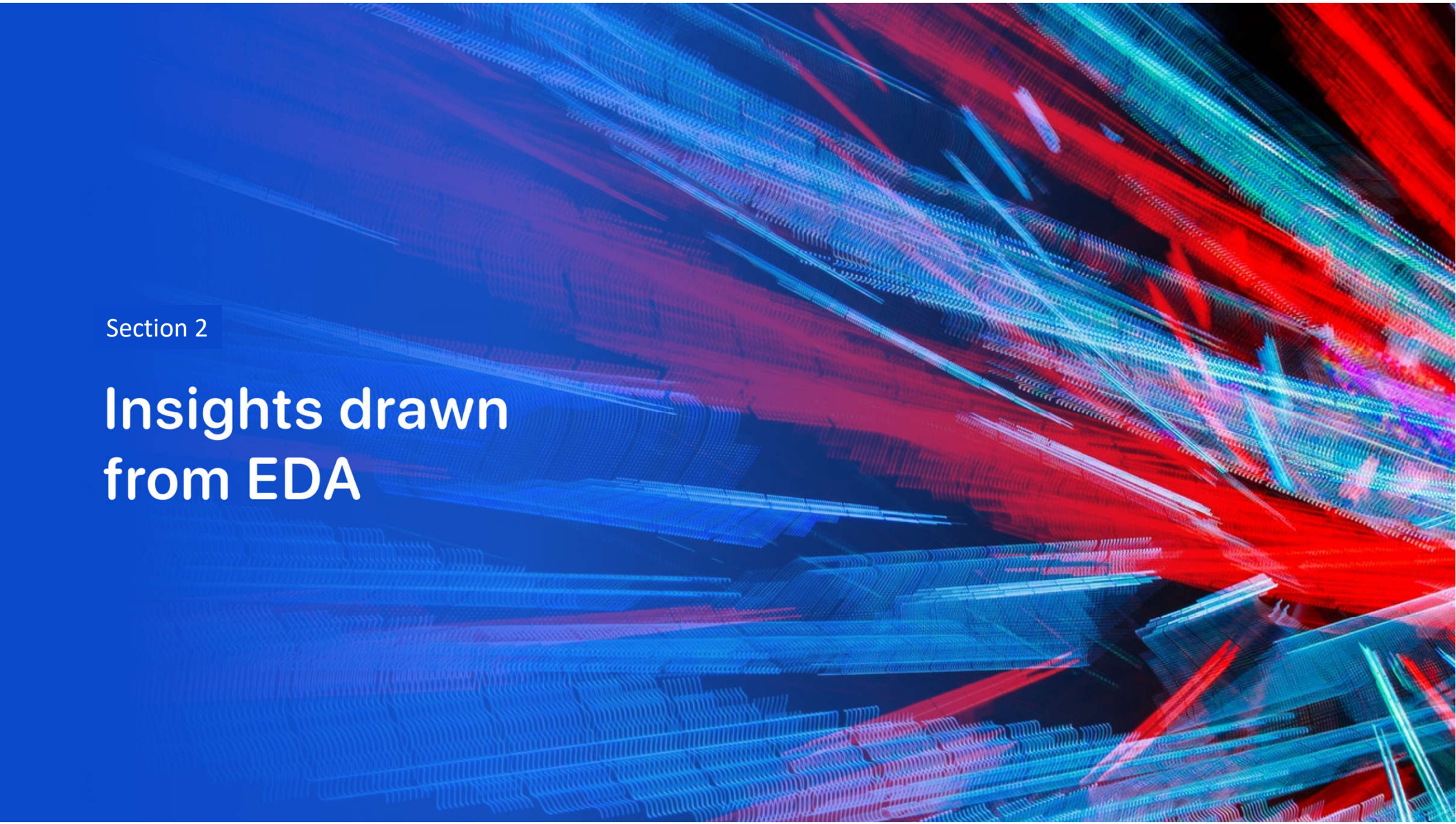
Link to GitHub Jupyter Notebook for Machine Learning Prediction

Predictive Analysis- Model Development Flowchart



Results

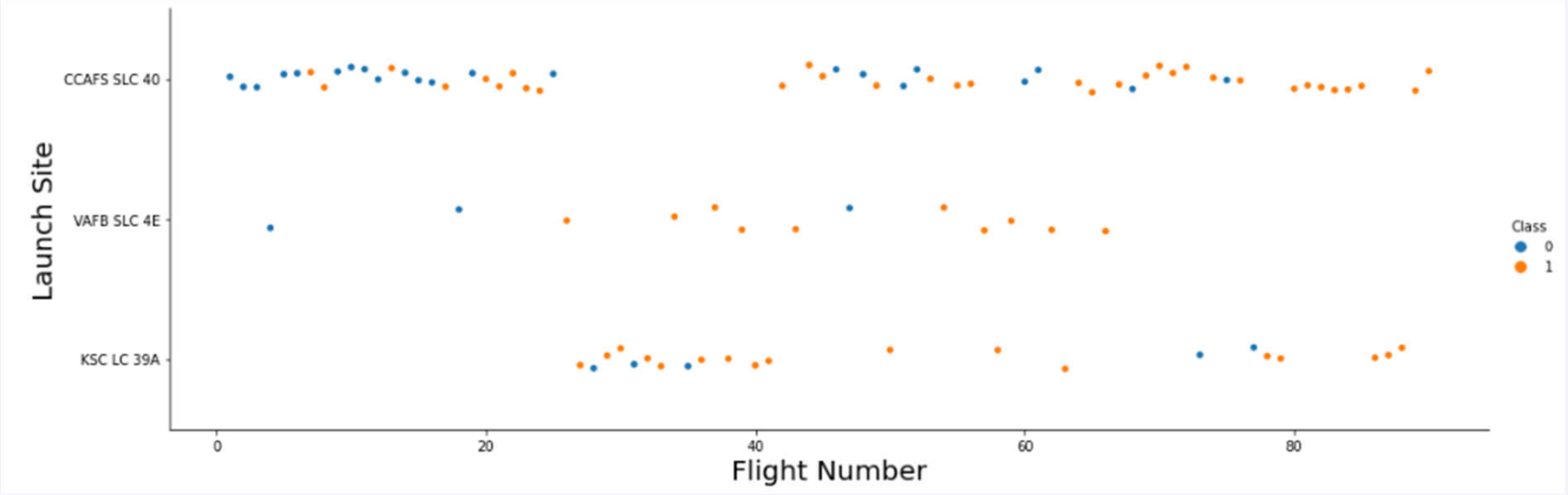
- SpaceX uses 4 launch sites.
- The average payload of the Falcon 9 booster is 2928 kg.
- The maximum payload of the Falcon 9 booster is 15,600 kg.
- The first lauch for SpaceX was in 2010, however the first successful booster landing was not achieved until 2013.
- Since 2013, SpaceX performance has steadily improved.
- Over the course of SpaceX operations, a 97% success rating has been achieved.
- Predictive analysis using a Classification Tree model results in 94% prediction accuracy.



Section 2

Insights drawn from EDA

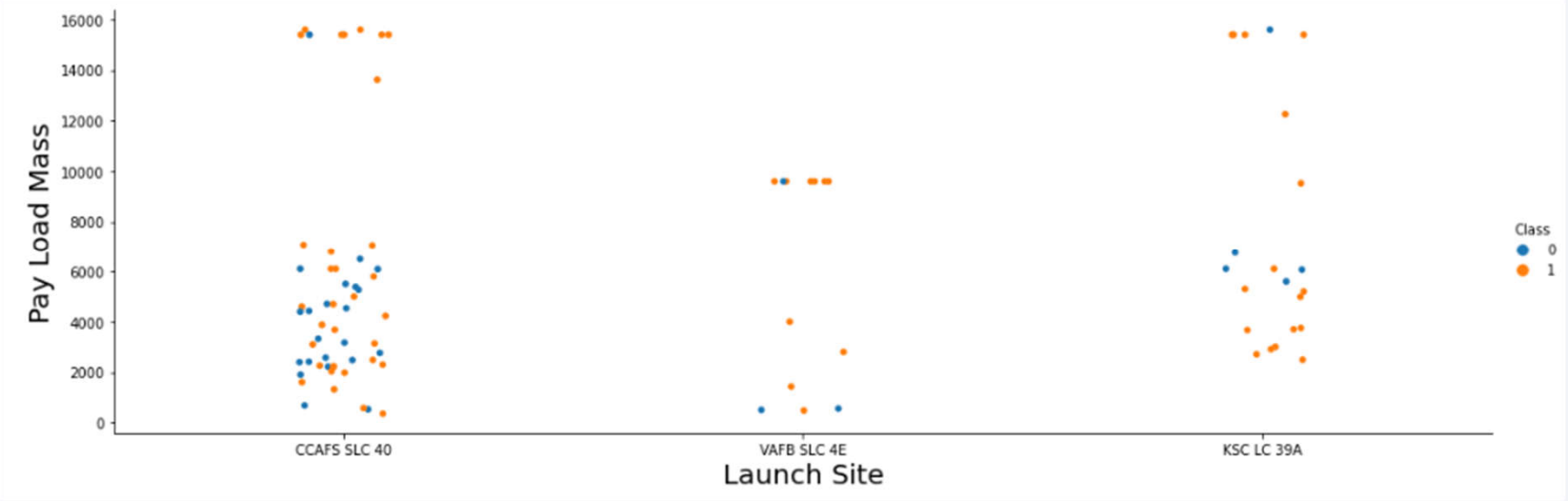
Flight Number vs. Launch Site



Class
0=Fail
1=Success

- It can be seen that as the number of flights increased, the rate of success seemed to improve.
- Cape Canaveral has the most numerous launches, with many failures at the start, but almost perfect performance after about 70 flights.
- Kennedy Space Center started launching SpaceX flights later, after about 30 total flights.

Payload vs. Launch Site

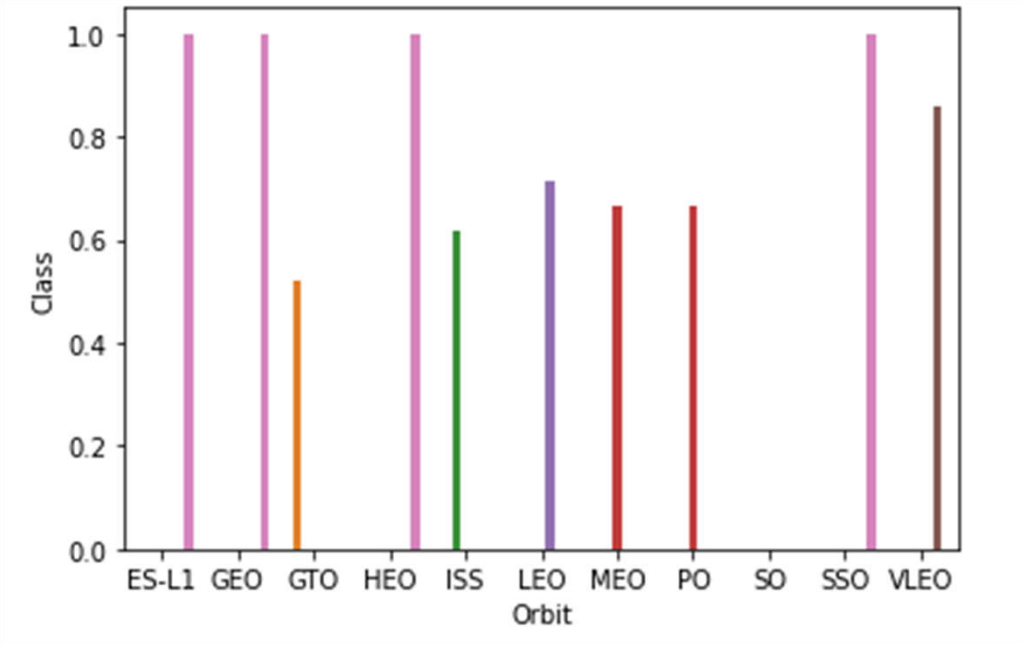


Class
0=Fail
1=Success

- It can be seen that Vandenberg AFB seems to have a top payload mass of 10,000 kg.
- The majority of flights across all launch sites are with payloads less than 8000kg.
- Cape Canaveral seems to have the largest number of failed missions, but this may relate to the fact that most of the earliest attempts (flights 1 through 30) were made here.

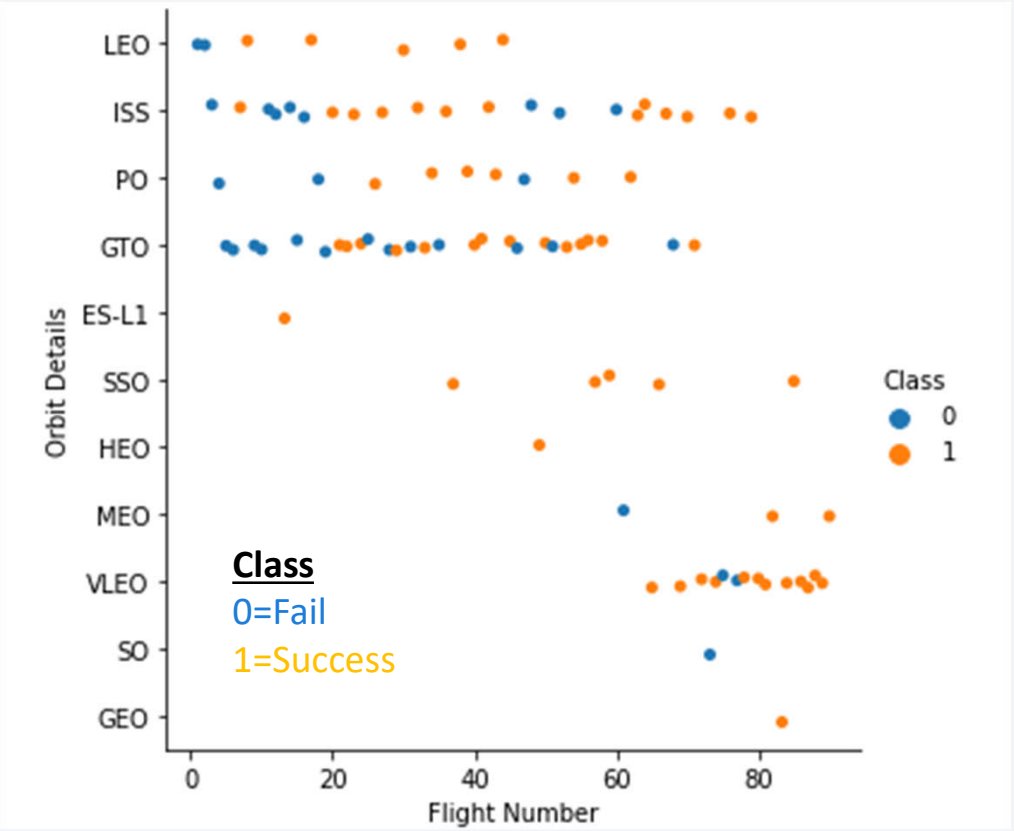
Success Rate vs. Orbit Type

- It can be seen low Earth Orbit (LEO) and Mid-Earth Orbit(MEO), as well as Geosynchronous Orbits (GTO) and Polar Orbits (PO) have relatively low success rates (50 to 60%).
- Counter-intuitively, the higher altitude orbits ES-L1, GEO, HEO, have relatively high success rates (100%).



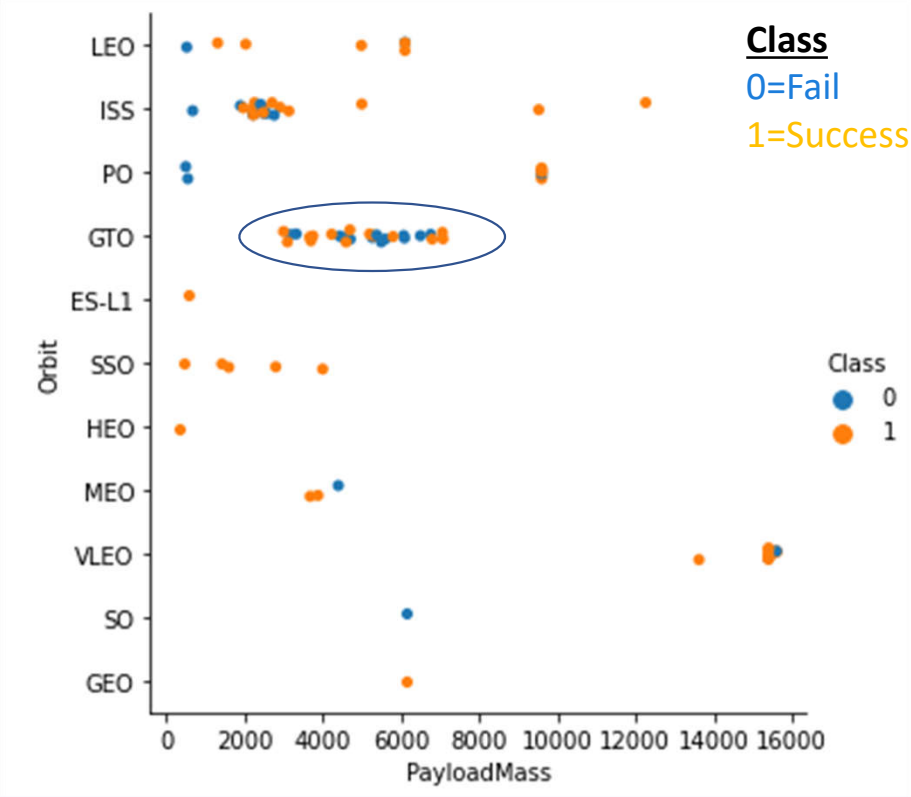
Flight Number vs. Orbit Type

- The previously observed effect of altitude on success rate may be more due to the disparate numbers of missions at each orbit.
- Lower altitude orbits tend to have large numbers of attempts (21-27 each).
- High altitude orbits, ES-L1, GEO and HEO, each have only 1 attempt each.



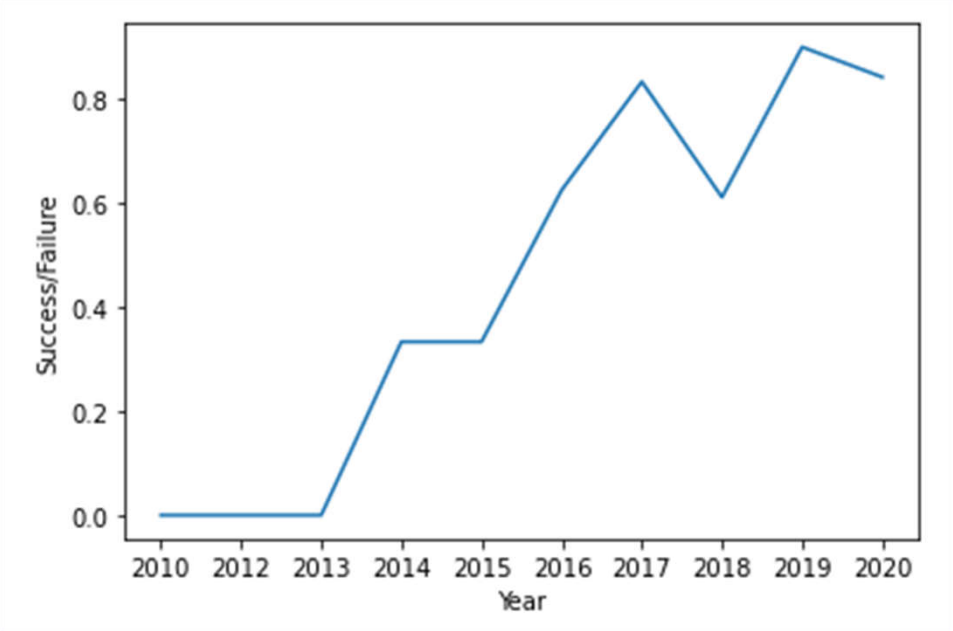
Payload vs. Orbit Type

- With heavy payloads, the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However, for GTO it is difficult to distinguish the effect of Orbit and Payload, as both positive landing rate and negative landing (unsuccessful mission) are equally represented.
- GTO has the highest number of attempted missions (27).



Launch Success Yearly Trend

- SpaceX has shown a consistent trend in increasing Success Rate during flight operations from 2010 through 2020.
- This could be due to a large number of factors:
  - Booster design improvements,
  - Increasing operator skill,
  - Better mission vetting and planning,
  - Improved weather forecasting.



All Launch Site Names

- Find the names of the unique launch sites

  %sql select distinct Launch_Site from SPACEX.TBL

- Launch Sites-

CCAFS LC-40	Cape Canaveral
CCAFS SLC-40	Cape Canaveral
KSC LC-39A	Kennedy Space Center
VAFB SLC-4E	Vandenberg Airforce Base

Launch Site Names Begin with 'KSC'

- Find 5 records where launch sites' names start with `KSC`

```
%sql select * from SPACEX.TBL where Launch_Site like 'KSC%' limit 5
```

- Five launches from Kennedy Space Center (KSC)

DATE	Time (UTC)	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing_outcome
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-03-16	06:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-01-05	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
2017-05-15	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

Total Payload Mass

- Calculate the total payload carried by boosters from NASA

```
%sql select sum(payload_mass__kg_) from SPACEX.TBL WHERE customer = 'NASA (CRS)'
```

- Total Payload Mass launched by NASA- 45,596 kg

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1

```
%sql select avg(payload_mass__kg_) from SPACEX.TBL WHERE booster_version = 'F9 v1.1'
```

- Average Falcon 9 Payload Mass- 2,928 kg

First Successful Landing Dates- Ground and Ship

- Find the dates of the first successful landing outcome on drone ship.

  %sql select min(DATE) from SPACEX.TBL WHERE LANDING_OUTCOME = 'Success (drone ship)'

  27 May 2016

- Find the dates of the first successful landing outcome on ground pad.

  %sql select min(DATE) from SPACEX.TBL WHERE LANDING_OUTCOME = 'Success (ground pad)'

  22 Dec 2015

The instructions for this slide were contradictory, so earliest dates of BOTH 'ground pad' and 'drone ship' landing are shown.

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
%sql select booster_version from SPACEX.TBL where Landing_Outcome = 'Success (drone ship)'\ and payload_mass__kg_ between 4000 and 6000
```

- Falcon 9 boosters with successful Drone Ship landings with Payload Mass from 4000 to 6000 kg-

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

  %sql select mission_outcome, count(mission_outcome) from SPACEX.TBL GROUP BY mission_outcome

- Present your query result with a short explanation here

mission_outcome	2
Failure (in flight)	2
Success	242
Success (payload status unclear)	3

SpaceX has an admirable 98.8% mission success rate. This makes SpaceX an attractive commercial option for launching orbital pay loads. 98.0% if “unclear payload status” is considered.

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

```
%sql select booster_version, payload_mass__kg_ from SPACEX.TBL\where
payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEX.TBL)
```

- Present your query result with a short explanation here:
  - A total of 12 Falcon 9 boosters have been launched carrying the maximum payload of 15,600 kg

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2017 Launch Records

- List the records which will display the month names, succesfull landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017

```
%sql SELECT TO_CHAR(TO_DATE(MONTH("DATE"), 'MM'), 'MONTH') AS MONTH_NAME, \
    LANDING_OUTCOME AS LANDING_OUTCOME, \
    BOOSTER_VERSION AS BOOSTER_VERSION, \
    LAUNCH_SITE AS LAUNCH_SITE \
    FROM SPACEX.TBL WHERE LANDING_OUTCOME = 'Success (ground pad)' AND "DATE" LIKE '%2017%'
```

- There were 6 successful landing outcomes for the Falcon 9 booster in 2017.
- Kennedy Space Center had 5 successful missions, Cape Canaveral had1.

month_name	landing_outcome	booster_version	launch_site
JANUARY	Success (ground pad)	F9 FT B1032.1	KSC LC-39A
FEBRUARY	Success (ground pad)	F9 FT B1031.1	KSC LC-39A
MARCH	Success (ground pad)	F9 FT B1035.1	KSC LC-39A
JULY	Success (ground pad)	F9 B4 B1040.1	KSC LC-39A
AUGUST	Success (ground pad)	F9 B4 B1039.1	KSC LC-39A
DECEMBER	Success (ground pad)	F9 FT B1035.2	CCAFS SLC-40

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

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

```
%sql select count(landing__outcome), landing__outcome from SPACEXTBL \
where DATE between '2010-06-04' and '2017-03-20' group by landing__outcome\
order by count(landing__outcome) desc
```

- During the period of 04 Jun 2010 and 20 Mar 2017, there were a total of 24 successful landings, equally distributed between ‘drone ship’ and ‘ground pad’ facilities.

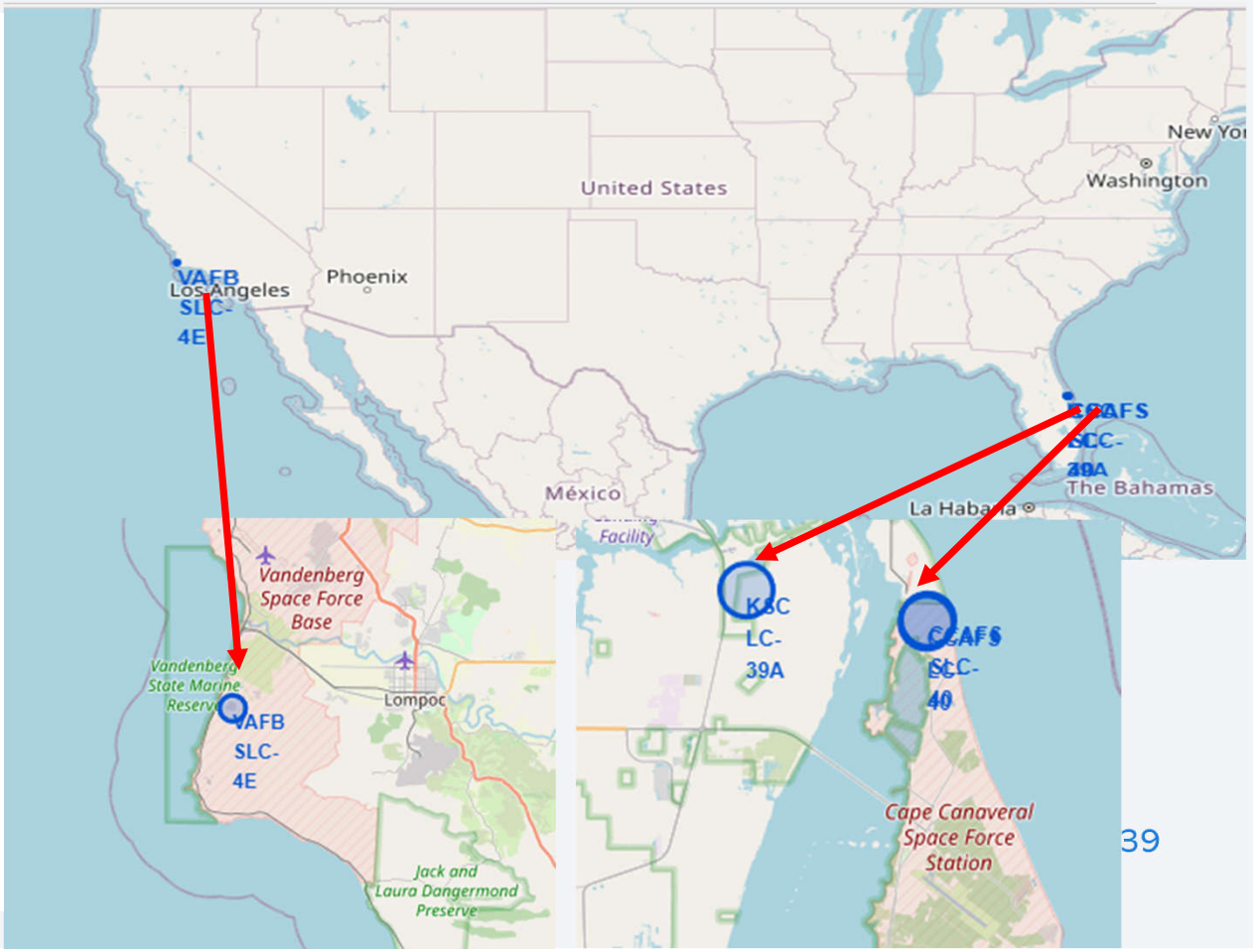
	landing_outcome
12	Success (drone ship)
12	Success (ground pad)

Section 3

Launch Sites Proximities Analysis

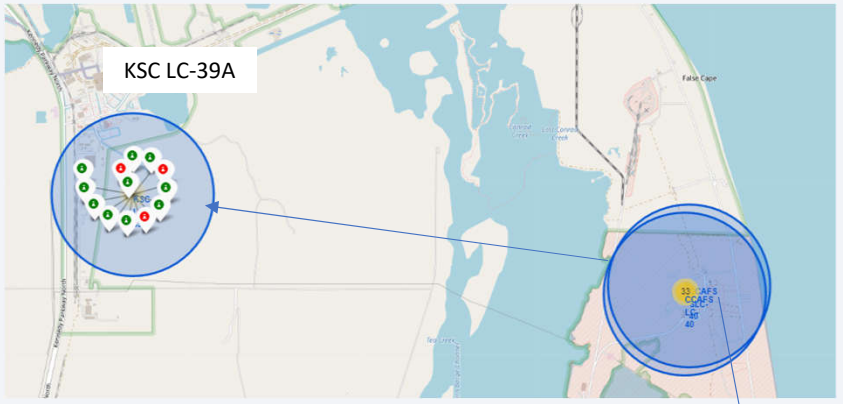
SpaceX Launch Sites

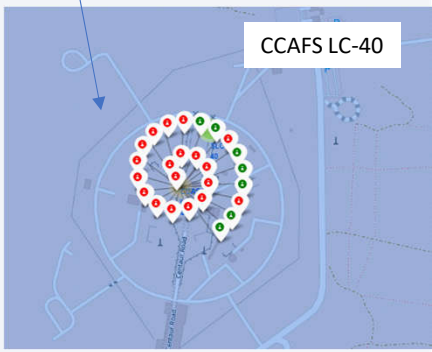
- The locations of SpaceX launch sites are positioned mostly toward the southern part of the US, closer to the Equator, to improve attaining orbital velocity more efficiently.
- In general, the launch sites are within a few (<1 to <3) KM of a coastline. This should enable Mission Control to preemptively detonate a rogue booster over water, as opposed to populated areas, in the case of a failed launch.
- Launch sites are kept a certain distance from major cities, the distance from KSC to Orlando, FL is 71.8KM, while the distance from VAFB to Lompoc, CA is 13.8 KM. This is proabably for 2 reasons, the risk of failed launches mentioned before, and also noise abatement for the populated areas during launches.

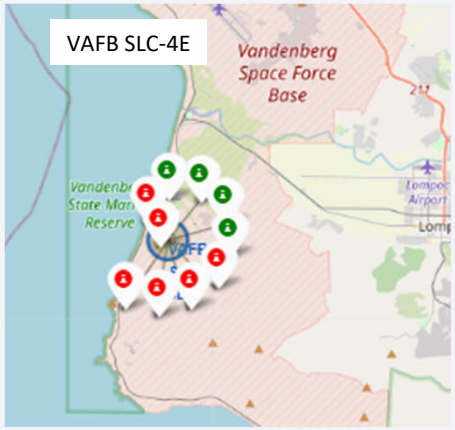


Launch Site Success/Fail

- Marker Clusters at Kennedy Space Center (KSC) , Cape Canaveral (CCAFS), and Vandenberg (VAFB), showing Success/Fail of launches
- It can be seen that Cape Canaveral has the most launch attempts, with more failures, probably associated with early test launches.
- Kennedy Space Center exhibits fewer launch failures, as it started later, benefitting from experience at CCAFS

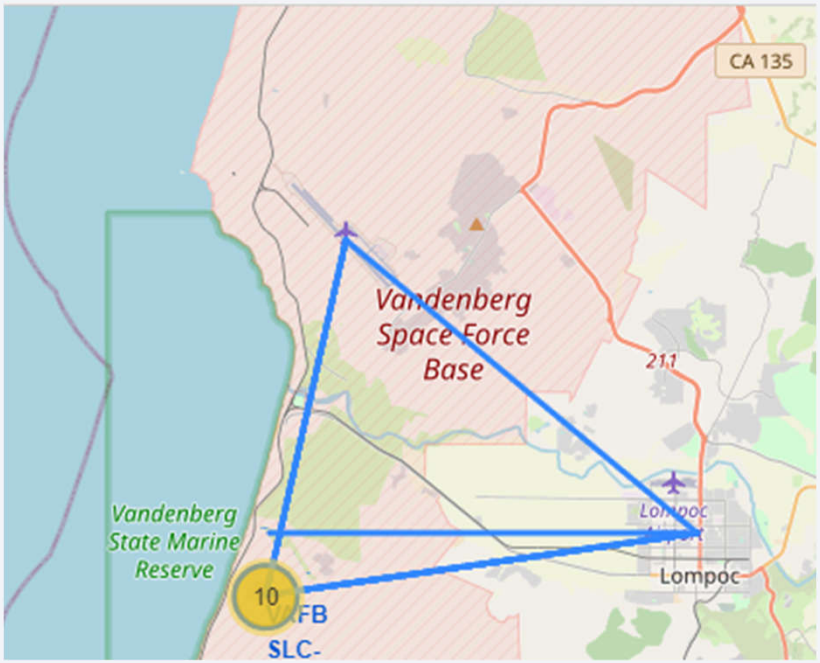


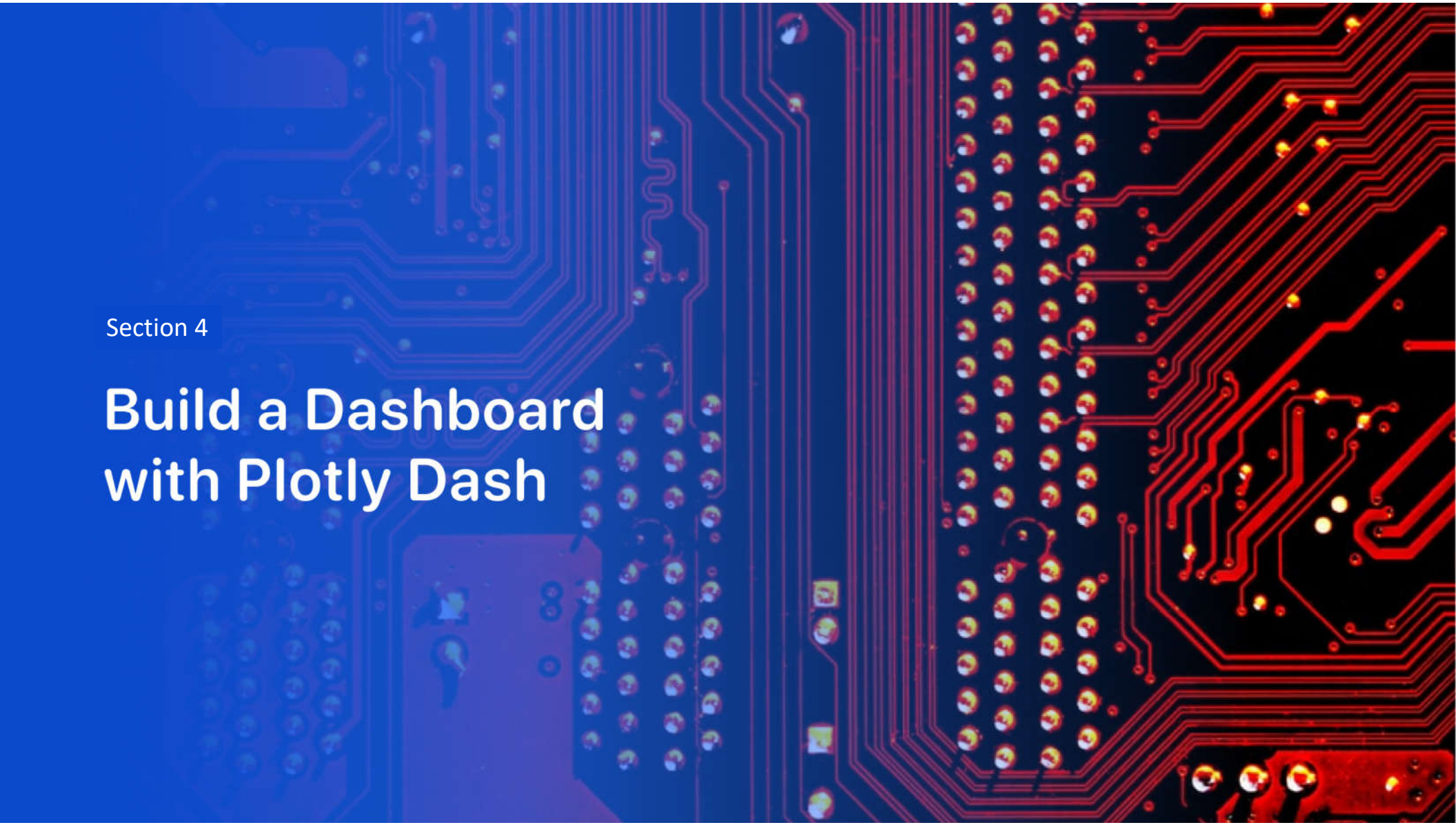




Measuring Distances from Launch Site SLC-4E

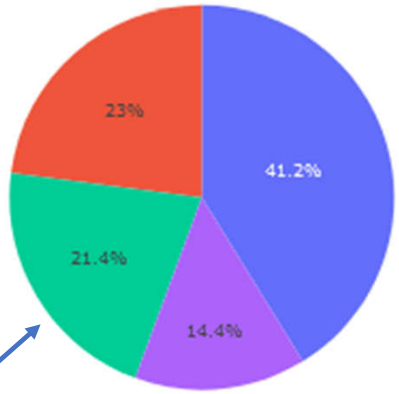
- 9.72 km Launch Site SLC-4E to VDAFB
- 13.76 km Launch Site SLC-4E to Lompoc, CA
- 13.75 km VDAFB to Lompoc, CA
- The distance from the launch site to the Space Force Base is critical, as all launch vehicles, fuel, payload will be sourced from the base to the launch facility.
- Distance from the base and launch facility to the nearest municipality may be important in terms of housing and support for personnel, as well as noise abatement during launch.

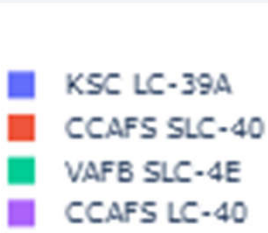


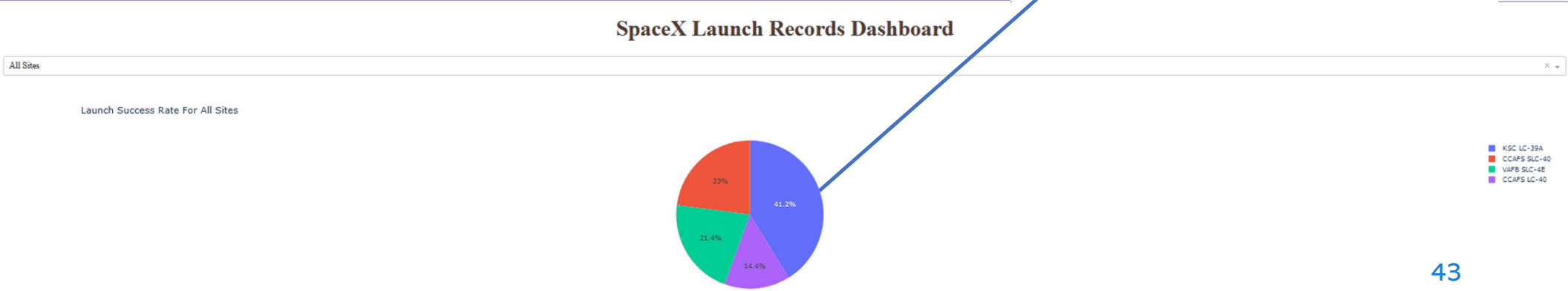

Section 4
Build a Dashboard
with Plotly Dash

Success Rate for All Sites

- The piechart for All Sites shows that Kennedy Space Center has the highest overall success rate.
- Cape Canaveral has the lowest success rate overall between CCAFS SLC-40 and LC-40.
- This may be because many of the early developmental launches were conducted from these sites, while Kennedy did not start launching until Flight 27 in 2017.

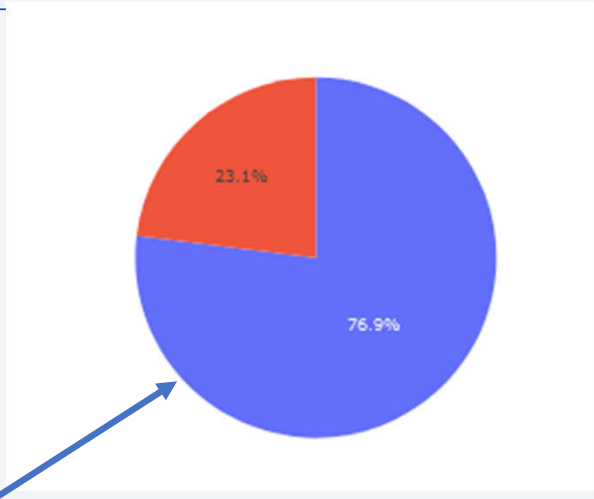


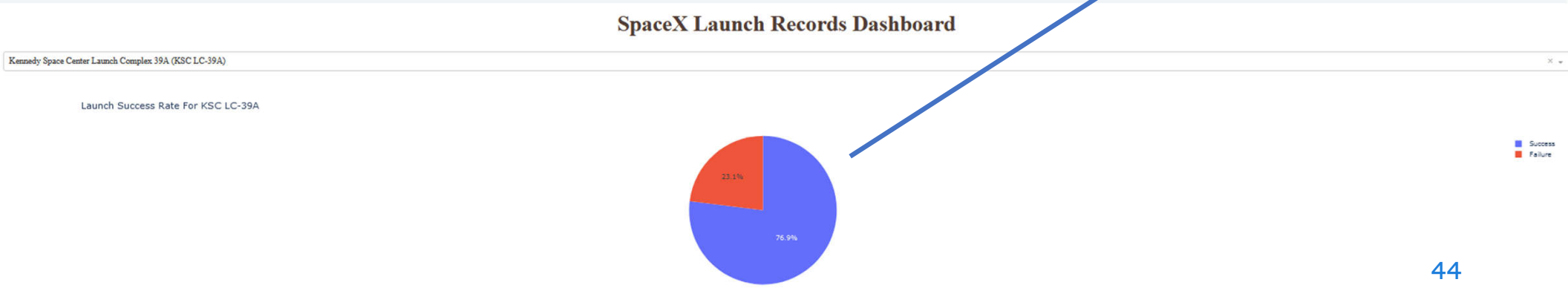




Success Rate- Kennedy Space Center

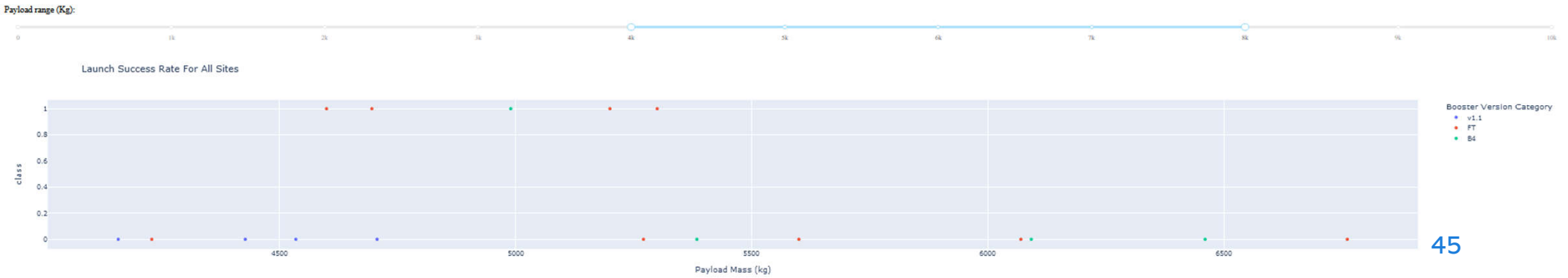
- The piechart for Kennedy Space Center has the highest overall success rate Of 76.9%5.
- Operations at KSC may have benefitted from early learnings at Cape Canaveral.





Success Rate- Payload and Booster Version

- The data in this chart show that most successful launches are for payloads in a moderate range, between 4500 and 5500 kg.
- Booster Version FT seem to provide the highest degree of reliability in this weight range.



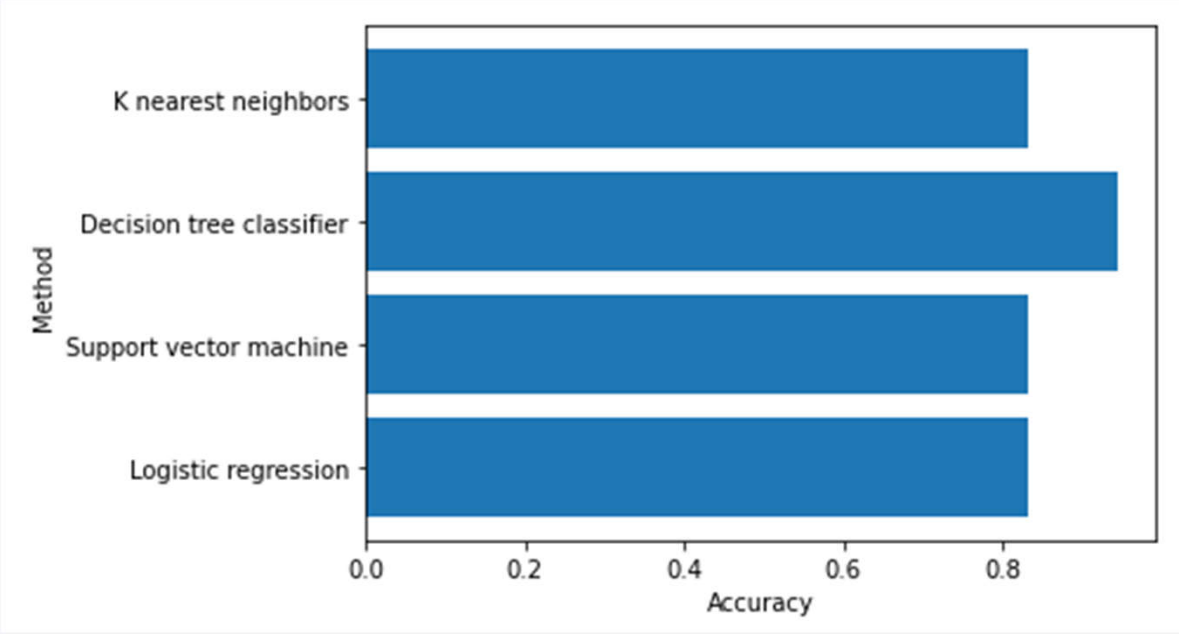
Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Visualize the built model accuracy for all built classification models, in a bar chart

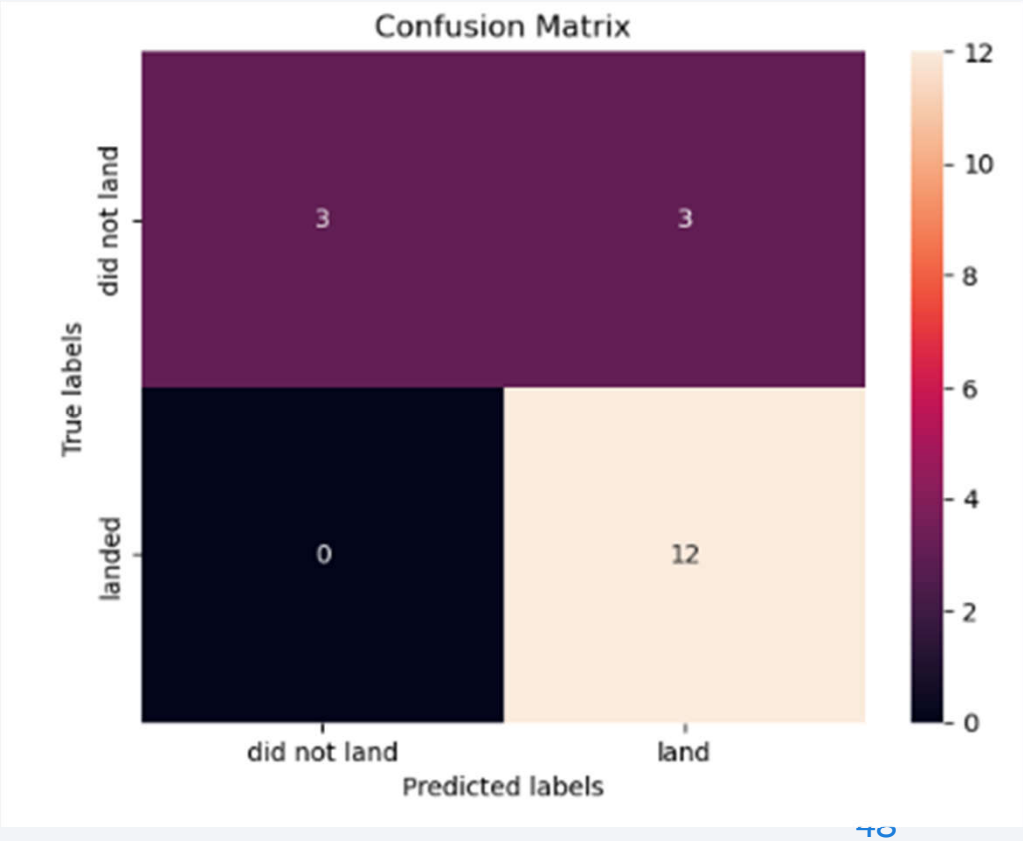
- The Decision Tree Classifier has the highest classification accuracy, 94%.



Model	accuracy
Logistic regression	0.833333
Support vector machine	0.833333
Decision tree classifier	0.944444
K nearest neighbors	0.833333

Confusion Matrix

- The confusion matrix shows that even with a 94% accuracy rate on the test data, there is still a bit of an issue with overprediction of a successful booster landing (3/15 = 20%).
- Performance of the model ought to be improved, perhaps with additional launch data.



Conclusions

- Different data sources were combined to produce a relatively successful predictive model for Falcon 9 booster landing.
- The best Launch Site is Kennedy Space Center KDC KC-39A
- For heavy payloads, the successful landing rates are high for Polar, LEO and ISS orbits.
- While SpaceX has compiled an impressive performance of 98% successful missions, this has been built up year over year after an experience of 3 years of flights before achieving a successful landing.
- The Decision Tree Classifier, while needing some improvement can be used to predict successful Falcon 9 booster landings.

Appendix

- Folium python files did not work on GitHub, I provided a link to Watson Studio version, as well as screenshots.
- For better visibility, a pdf of the data treatment and model development flowcharts is provided:

Foxit
PhantomPDF PDF Document

- Link to GitHub Repository:

https://github.com/JTKE10901/Capstone-SpaceX-Project.git

Thank you!