IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Prithvi Raj
September 20, 2024

TABLE OF CONTENTS

EXECUTIVE SUMMARY

- SpaceY is a new commercial rocket launch provider who wants to bid against SpaceX.
- SpaceX advertises launch services starting at $62 million for missions that allow some fuel to be reserved for landing the 1st stage rocket booster, so that it can be reused.
- SpaceX public statements indicate a 1st stage Falcon 9 booster to cost upwards of $15 million to build without including R&D cost recoupment or profit margin.
- Given mission parameters such as payload mass and desired orbit, the models produced in this report were able to predict the first stage rocket booster landing successfully with an accuracy level of 83.3%.
- As a result, SpaceY will be able to make more informed bids against SpaceX by using 1st stage landing predictions as a proxy for the cost of a launch.

INTRODUCTION: BACKGROUND

- This report has been prepared as part of the Applied Data Science Capstone course.*
- In this capstone, I take the role of a data scientist working for a new rocket company called SpaceY.
- With the help of the data science findings and models in this report, SpaceY will be able to make more informed bids against SpaceX for a rocket launch.

* 10th course in the IBM Data Science Professional Certification

INTRODUCTION: BUSINESS PROBLEM

Compilation of early attempts at propulsive landing prior to the first success in 2016

- SpaceX advertises Falcon 9 rocket launches with a cost of 62 million dollars when the first stage of their rockets can be reused.
- The first stage is estimated to cost upwards of 15 million to build without including R&D cost recoupment or profit margin.
- Sometimes SpaceX will sacrifice the first stage due to mission parameters such as payload, orbit, and customer.
- Therefore this report aims to accurately predict the likelihood of the first stage rocket landing successfully as a proxy for the cost of a launch.

Section 1

Methodology

METHODOLOGY

For this report the data science methodology used can be outlined as such:

1. Data collection
2. Data wrangling
3. Exploratory data analysis
4. Data visualization
5. Model development
6. Reporting results to stakeholders

METHODOLOGY


Past launches

2010 to 2013

What the first page of launch data looked like on Wikipedia prior to web scraping

- Data Collection
  - API
    - Acquired historical launch data from Open Source REST API for SpaceX
      - Requested and parsed the SpaceX launch data using the GET request
      - Filtered the dataframe to only include Falcon 9 launches
      - Replaced missing payload mass values from classified missions with mean
  - Web Scraping
    - Acquired historical launch data from Wikipedia page 'List of Falcon 9 and Falcon Heavy Launches'
      - Requested the Falcon9 Launch Wiki page from its Wikipedia URL
      - Extracted all column/variable names from the HTML table header
      - Parsed the table and converted it into a Pandas data frame

Note: Falcon 9 launch dataset was limited to launches before December 7, 2020 per instructions.

METHODOLOGY

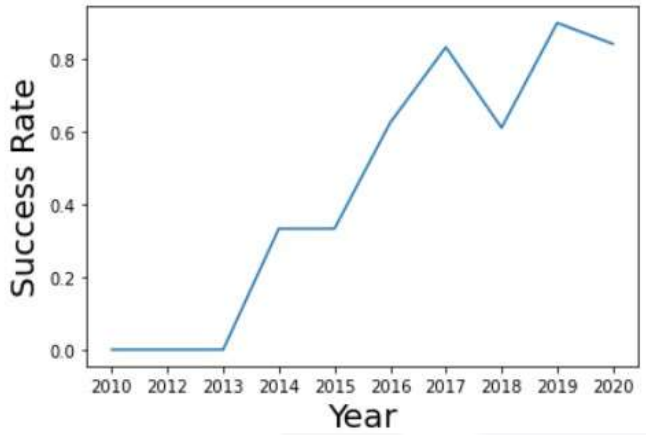
Landing Outcomes
sample size = 90
□= Class 0
□= Class 1

True ASDS	41
None None	19
True RTLS	14
False ASDS	6
True Ocean	5
None ASDS	2
False Ocean	2
False RTLS	1

- Data Wrangling
  - Explored data to determine the label for training supervised models
    - Calculated the number of launches on each site
    - Calculated the number and occurrence of each orbit
    - Calculated the number and occurrence of mission outcome per orbit type
  - Created a landing outcome training label from 'Outcome' column
    - Training label: 'Class'
    - Class = 0; first stage booster did not land successfully
      - None None; not attempted
      - None ASDS; unable to be attempted due to launch failure
      - False ASDS; drone ship landing failed
      - False Ocean; ocean landing failed
      - False RTLS; ground pad landing failed
    - Class = 1; first stage booster landed successfully
      - True ASDS; drone ship landing succeeded
      - True RTLS; ground pad landing succeeded
      - True Ocean; ocean landing succeeded

METHODOLOGY

```
In [11]: # Plot a line chart with x axis to be the extracted year a
         nd y axis to be the success rate
         df1=pd.DataFrame(Extract_year(df['Date']),columns =['year
         '])
         df1['Class']=df['Class']
         sns.lineplot(data=df1, x=np.unique(Extract_year(df['Date
         ']))), y=df1.groupby('year')['Class'].mean())
         plt.xlabel("Year", fontsize=20)
         plt.ylabel("Success Rate", fontsize=20)
         plt.show()
```

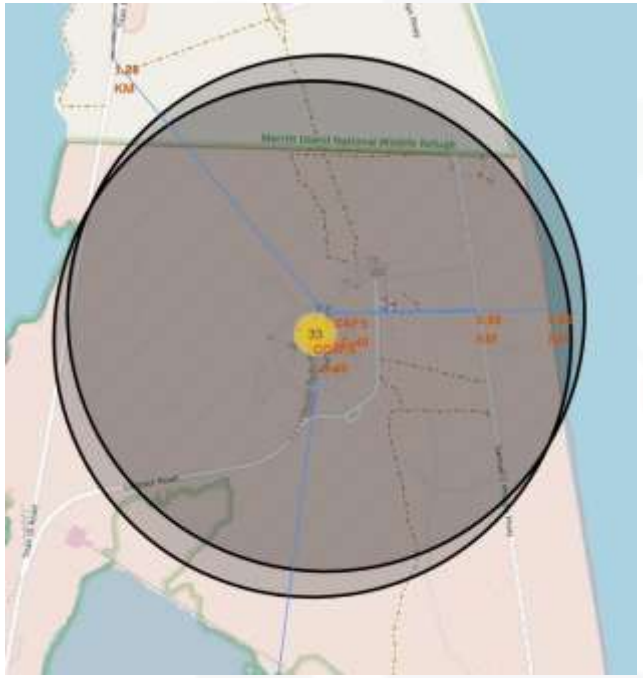


screenshot of Year vs. Success rate* plot
*(of 1st stage booster landing)

- Exploratory Data Analysis (EDA)
  - EDA with SQL
    - Loaded data into an IBM DB2 instance
    - Ran SQL queries to display and list information about
      - Launch sites
      - Payload masses
      - Booster versions
      - Mission outcomes
      - Booster landings
  - EDA with visualization
    - Read the dataset into a Pandas dataframe
    - Used Matplotlib and Seaborn visualization libraries to plot
      - FlightNumber x PayloadMass †
      - FlightNumber x LaunchSite †
      - Payload x LaunchSite †
      - Orbit type x Success rate
      - FlightNumber x Orbit type †
      - Payload x Orbit type †
      - Year x Success rate

† = with Class overlayed (1st stage booster landing outcome)

METHODOLOGY

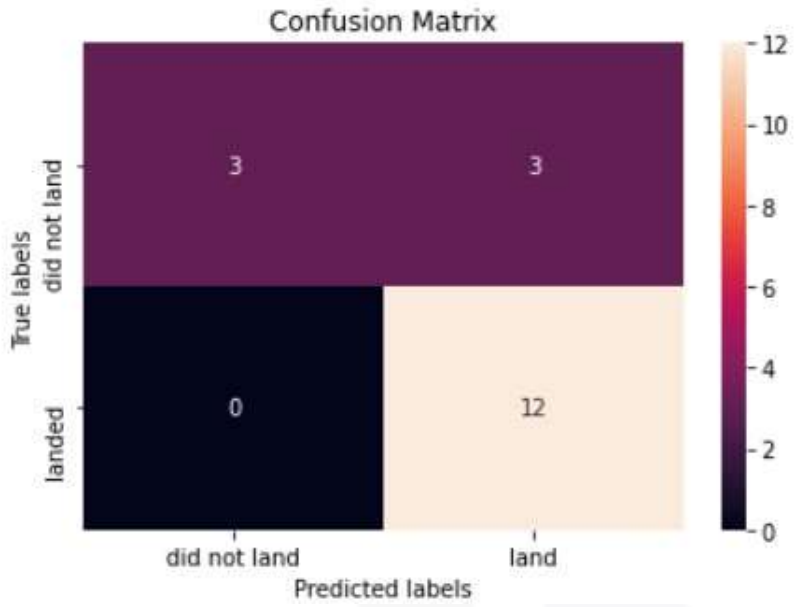


Screenshot of interactive Folium map showing proximity from CCAFS SLC-40 launch site to nearby railway, highway, and coastline

- Data Visualization
  - Launch Sites Location Analysis
    - Used Python interactive mapping library called Folium
    - Marked all launch sites on a map
    - Marked the successful/failed launches for each site on map
    - Calculated the distances between a launch site to its proximities
      - Railways
      - Highways
      - Coastlines
      - Cities
  - Launch Records Dashboard
    - Used Python interactive dashboarding library called Plotly Dash to enable stakeholders to explore and manipulate data in an interactive and real-time way
    - Pie chart showing success rate
      - Color coded by launch site
    - Scatter chart showing payload mass vs. landing outcome
      - Color coded by booster version
      - With range slider for limiting payload amount
    - Drop-down menu to choose between all sites and individual launch sites
    - Deployed to Heroku static web app hosting service

      https://ibm-applied-data-science-capst.herokuapp.com/

METHODOLOGY



Confusion matrix of logistic regression model, showing 15 correct predictions and 3 false positives

- Predictive Analysis (Model development)
  - Imported libraries and defined function to create confusion matrix
    - Pandas
    - Numpy
    - Matplotlib
    - Seaborn
    - Sklearn
  - Loaded the dataframe created during data collection
  - Created a column for our training label 'Class' created during data wrangling
  - Standardized the data
  - Split the data into training data and test data
  - Fit the training data to various model types
    - Logistic Regression
    - Support Vector Machine
    - Decision Tree Classifier
    - K Nearest Neighbors Classifier
  - Used a cross-validated grid-search over a variety of hyperparameters to select the best ones for each model
    - Enabled by Scikit-learn library function GridSearchCV
  - Evaluated accuracy of each model using test data to select the best model

RESULTS

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

RESULTS: EDA WITH SQL

The team at SpaceY had some very specific questions to answer with SQL:

- What launch sites has SpaceX used?
  - CCAFS LC-40
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E
- Examine launch site and date records where launch sites begin with the string 'CCA', do they overlap?
  - Last launch from CCAFS LC-40 was 2016-08-14
  - First launch from CCAFS SLC-40 was 2017-12-15
  - Wikipedia confirms Cape Canaveral Space Launch Complex 40 was renamed in 2017
- Display the total payload mass carried by boosters launched by NASA (CRS)
  - 45,596 KG, total
- Display average payload mass carried by booster version F9 v1.1
  - 340 KG, average
- List the date when the first successful landing outcome in ground pad was achieved.
  - 2015-12-22, more than 5 years after the first Falcon 9 launch on 2010-06-04

RESULTS: EDA WITH SQL (CONTINUED)

- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - F9 FT B1021.1
  - F9 FT B1023.1
  - F9 FT B1029.2
  - F9 FT B1038.1
  - F9 B4 B1042.1
  - F9 B4 B1045.1
  - F9 B5 B1046.1

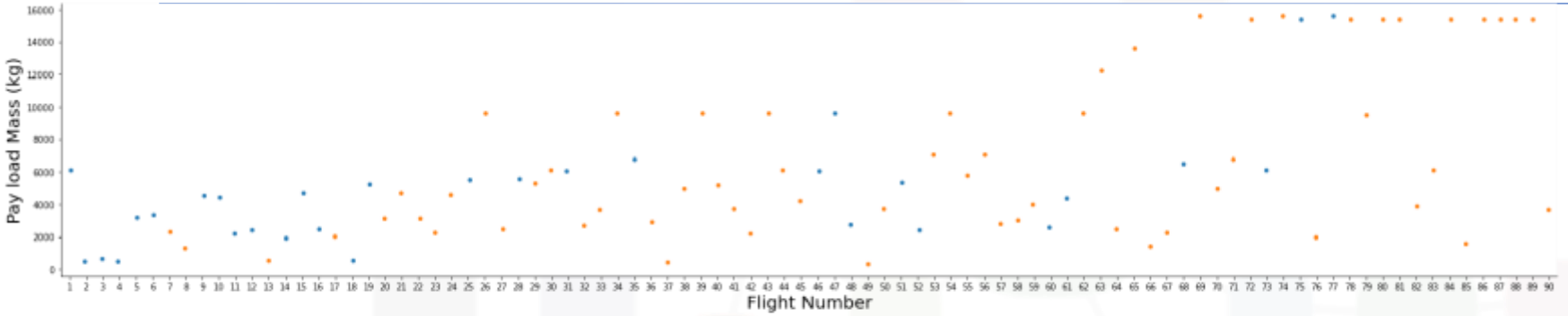
- Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order
  - 10 - No attempt5 - Failure (drone ship)
  - 5 - Success (drone ship)
  - 3 - Controlled (ocean)
  - 3 - Success (ground pad)
  - 2 - Failure (parachute)
  - 2 - Uncontrolled (ocean)
  - 1 - Precluded (drone ship)

- List the names of the booster_versions which have carried the maximum payload mass.
  - 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

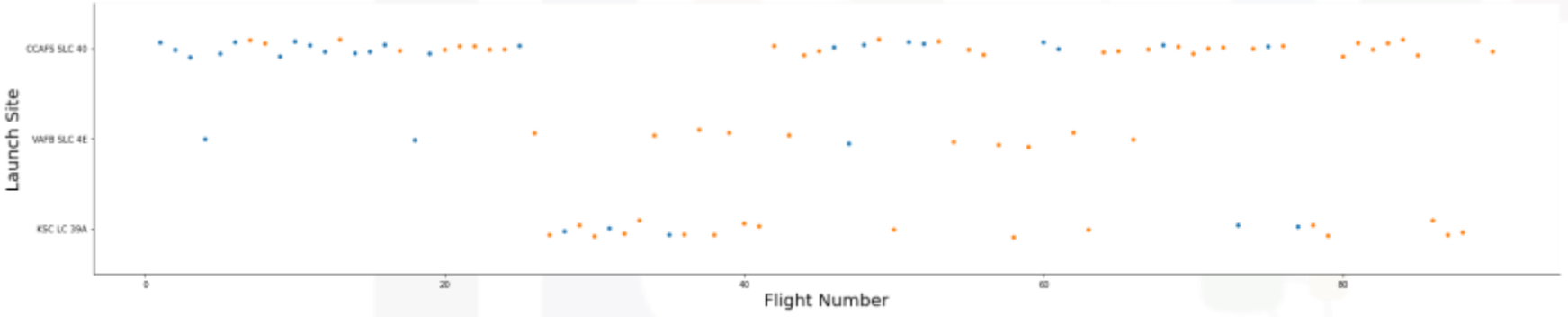
- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - Failure (drone ship) F9 v1.1 B1012 CCAFS LC-40
  - Failure (drone ship) F9 v1.1 B1015 CCAFS LC-40

- List the total number of successful and failure mission outcomes
  - 1 - Failure (in flight)
  - 99 - Success
  - 1 - Success (payload status unclear)

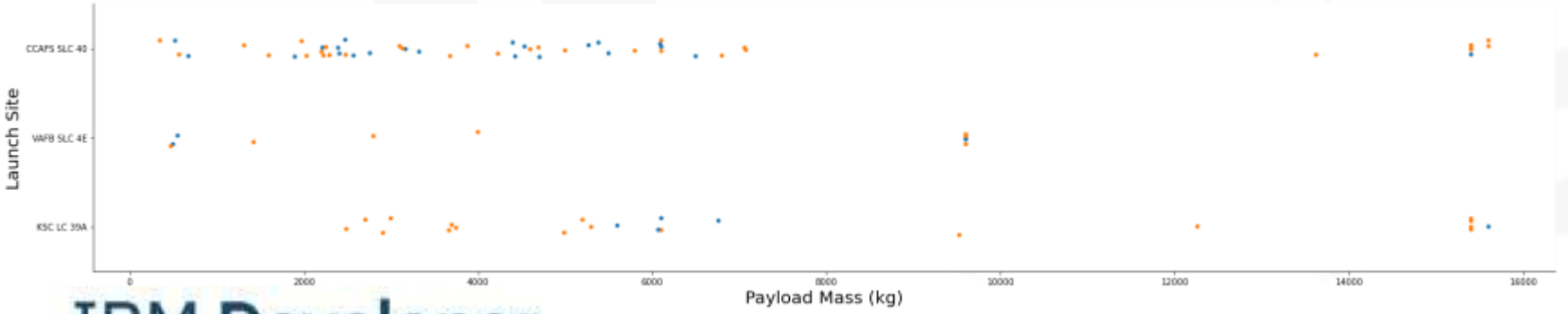
RESULTS: EDA WITH VISUALIZATION



FlightNumber x PayloadMass,
1st stage landing success positively correlated with continuous launch attempts, while negatively correlated with payload mass

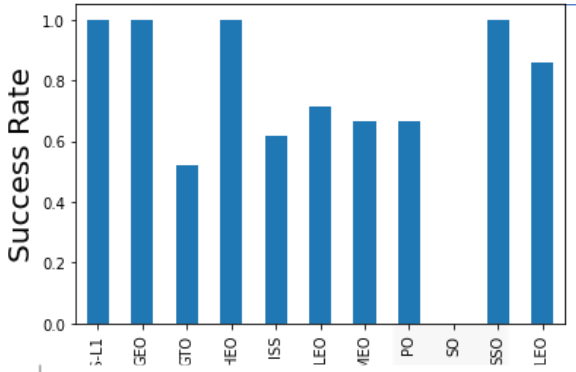


FlightNumber x LaunchSite,
CCAFS SLC 40 appears to have been where most of the early 1st stage landing failures took place

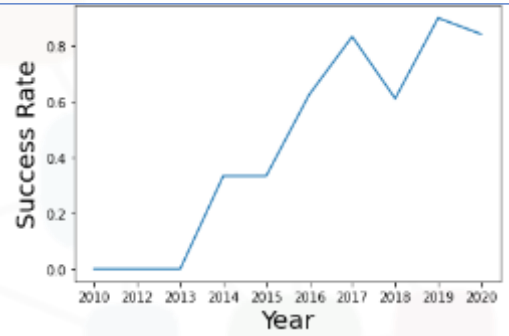


PayloadMass x LaunchSite,
CCAFS SLC 40 and KSC LC 39A appear to be favored for heavier payloads

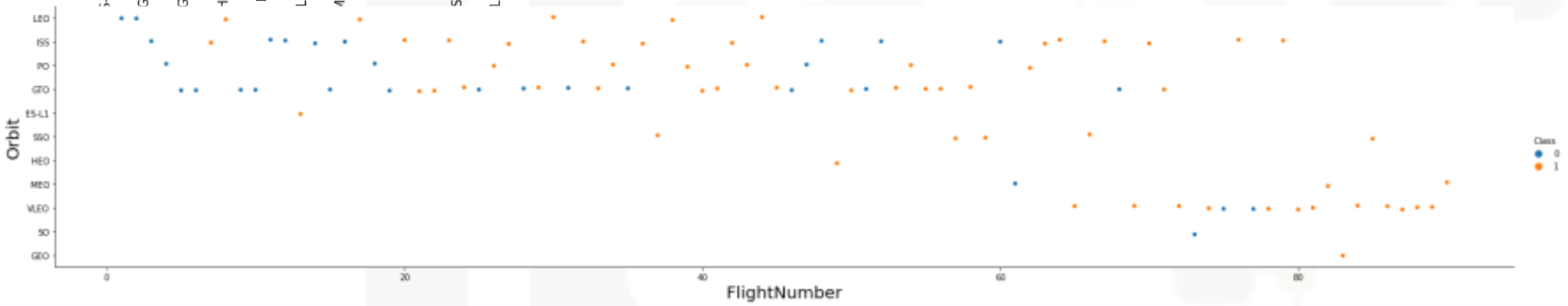
RESULTS: EDA WITH VISUALIZATION (CONTINUED)



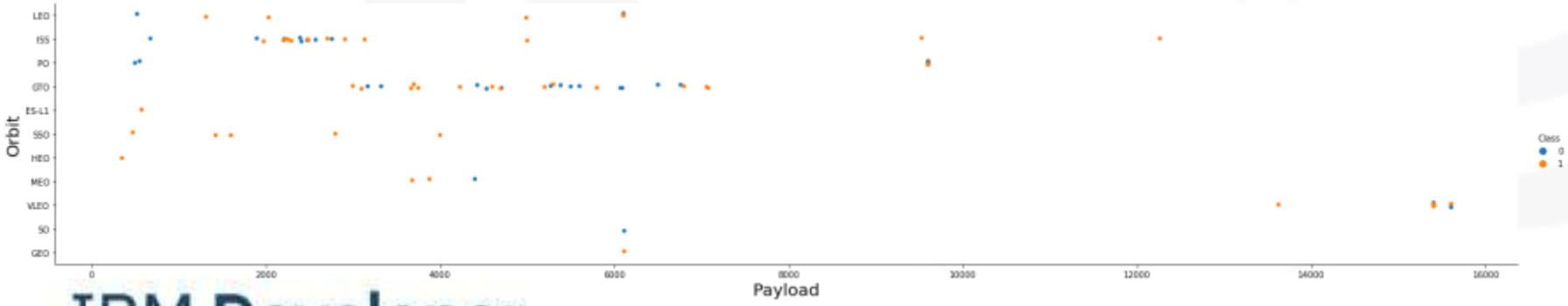
Orbit type x Success rate,
All orbit types except 'SO' have had successful 1st stage landings



Year x Success rate,
success rate trending positively on a yearly basis since 2013



FlightNumber x Orbit type,
flight number positively correlated with 1st stage recovery for all orbit types



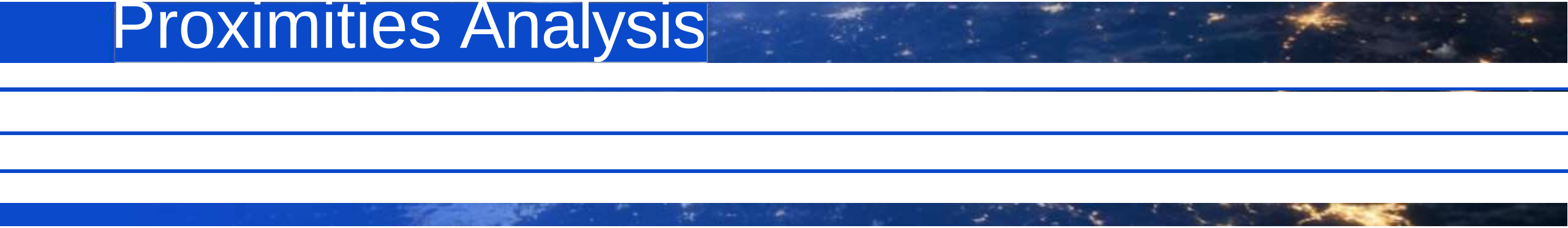
PayloadMass x Orbit type,
heavier payloads have a negative influence on GTO orbits and positive influence on ISS orbits

IBM Developer

SKILLS NETWORK

Section 4

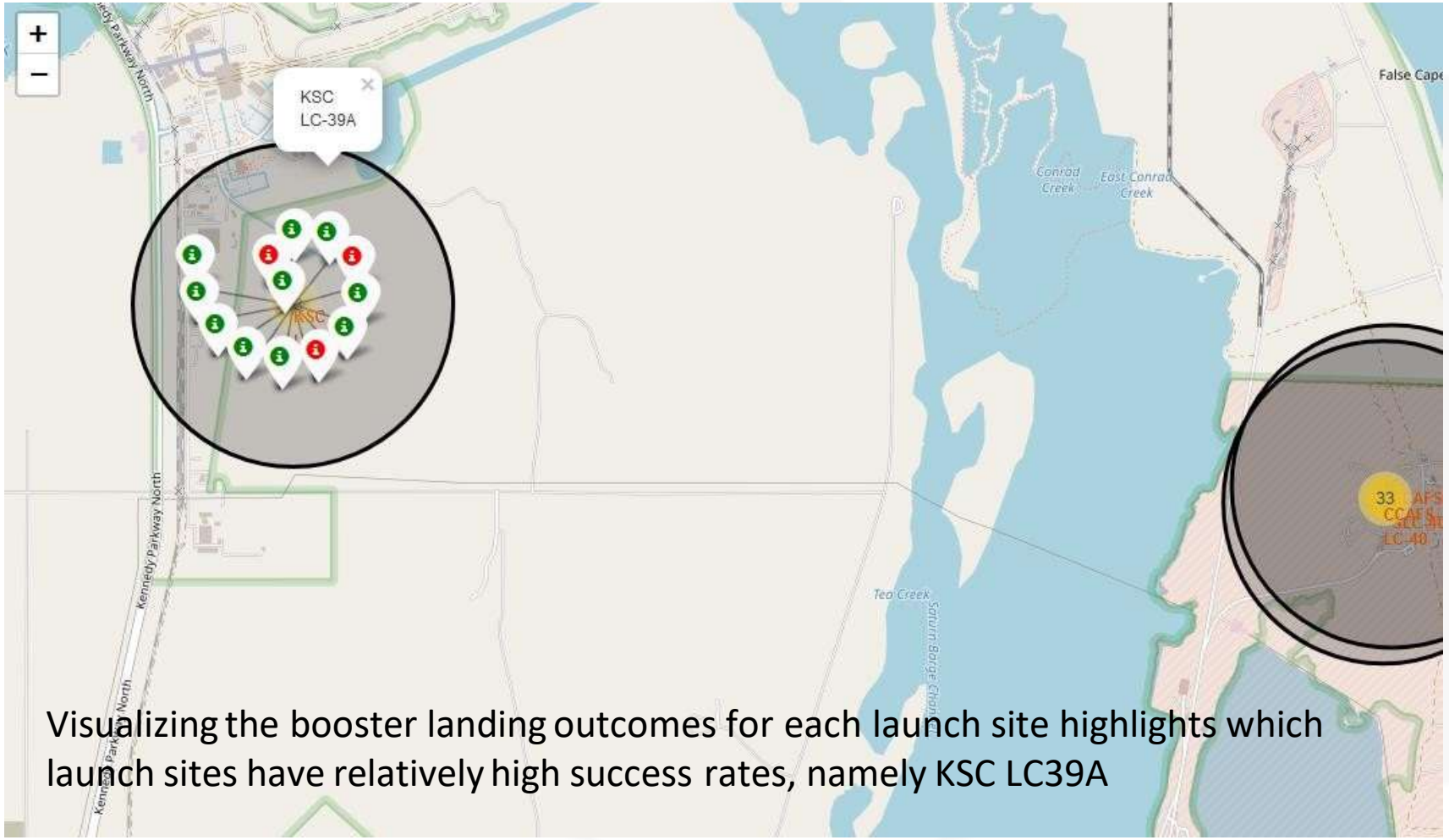
Launch Sites Proximities Analysis



RESULTS: LAUNCH SITE LOCATION ANALYSIS

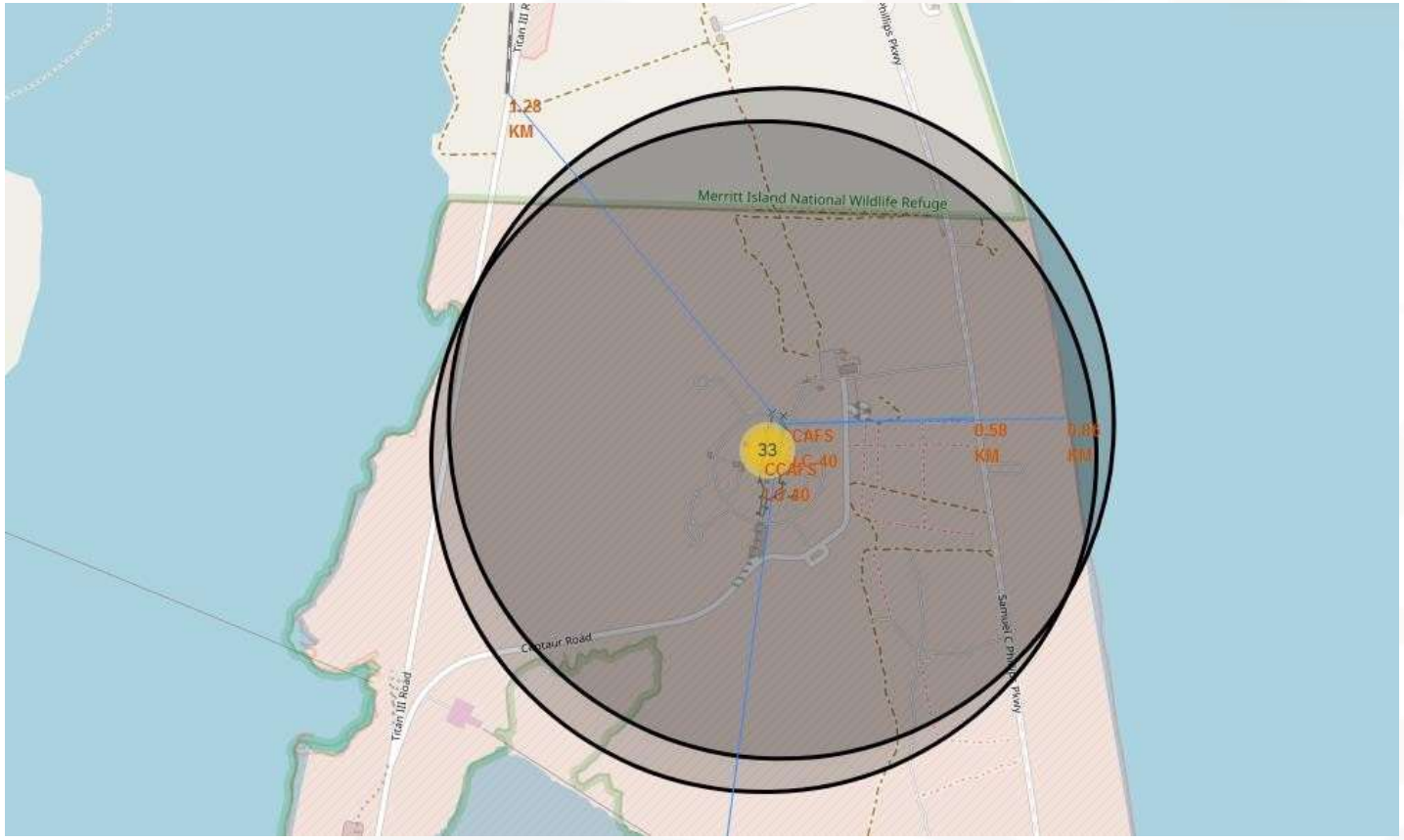
Visualizing the launch sites on a map highlights the importance of launch site proximity to the coast and equator:

RESULTS: LAUNCH SITE LOCATION ANALYSIS (CONTINUED)



Visualizing the booster landing outcomes for each launch site highlights which launch sites have relatively high success rates, namely KSC LC39A

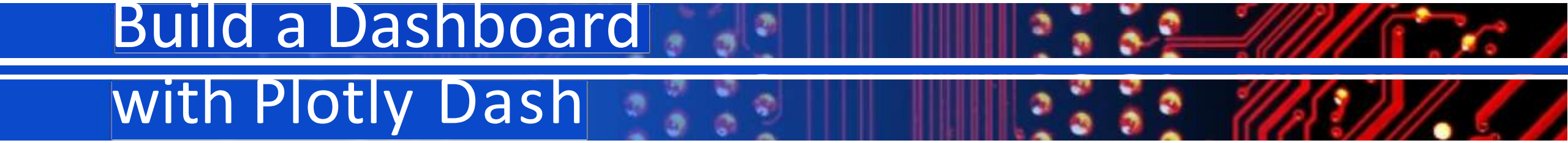
RESULTS: LAUNCH SITE LOCATION ANALYSIS (CONTINUED)



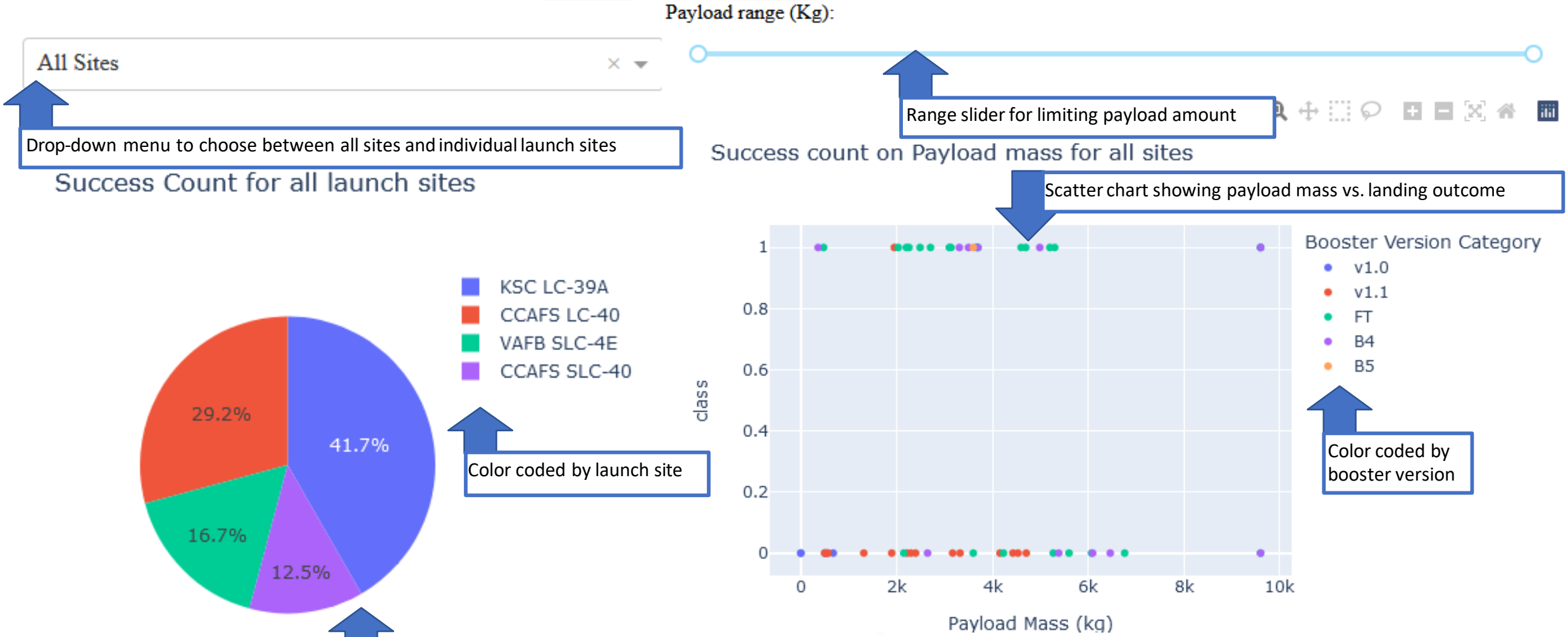
- Visualizing the railway, highway, coastline, and city proximities for each launch site allows us to see how close each is, for example:
- Proximities for CCAFS SLC-40:
  - railway: 1.28 km
    - transporting heavy cargo
  - highway: 0.58 km
    - transporting personel and equipment
  - coastline: 0.86 km
    - optionality to abort launch and attempt water landing
    - minimizing risk from falling debris
  - city: 51.43 km
    - minimizing danger to population dense areas.

Section 5

Build a Dashboard with Plotly Dash



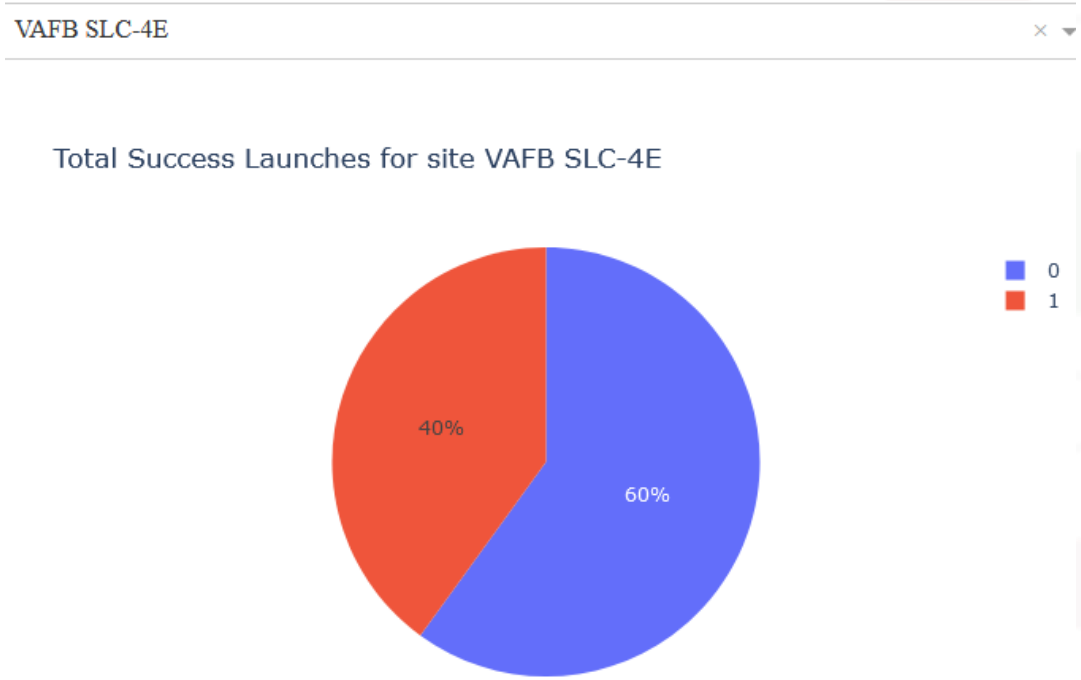
RESULTS: LAUNCH RECORDS DASHBOARD



IBM Developer

SKILLS NETWORK

RESULTS: LAUNCH RECORDS DASHBOARD (CONTINUED)



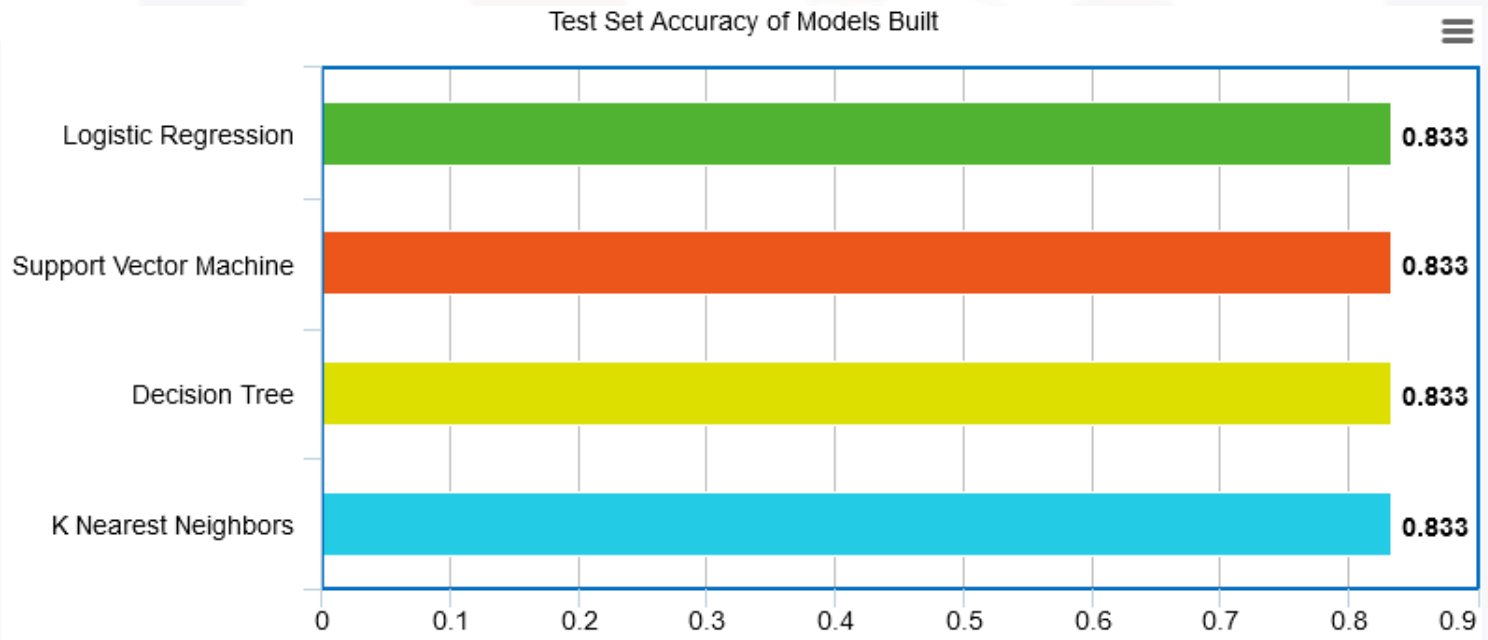
Example dashboard view:
Booster landing sucess rate for VAFB SLC-4E

- Explore the dashboard yourself:
  - https://ibm-applied-data-science-capst.herokuapp.com/
  - Enabling stakeholders to explore and manipulate the data in an interactive and real-time way
- Dashboard observations:
  - FAFB SLC-4E had the heaviest successful booster landing success
  - KSC LC-39A has the highest booster landing success rate
  - Payloads < 5,300 kg had the highest booster landing success rate
  - Payloads > 5,300 kg had the lowest booster landing success rate

Section 6

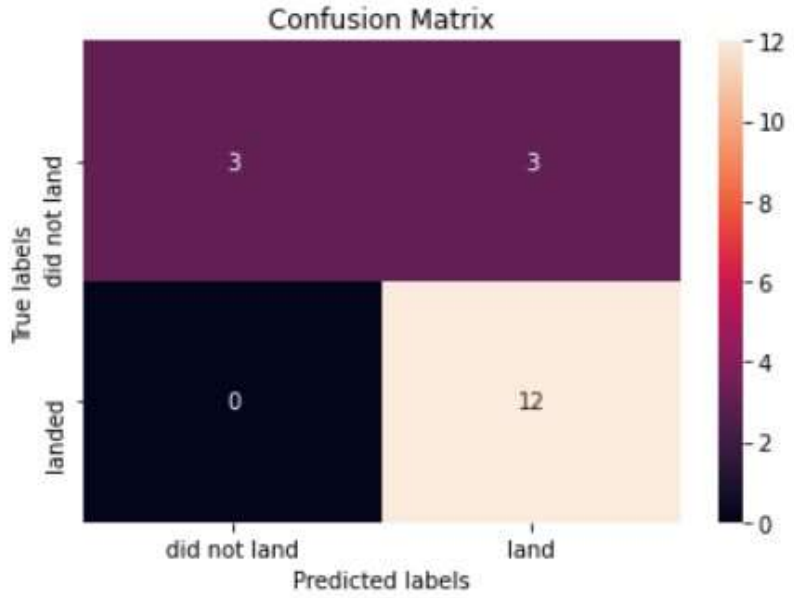
Predictive Analysis (Classification)

RESULTS: CLASSIFICATION ACCURACY



- Each of the four models built came back with the same accuracy score, 83.33%

RESULTS: CONFUSION MATRIX



- The confusion matrices of the best performing models (4-way-tie) are the same
- The major problem is false positives as evidenced by the models incorrectly predicting the 1st stage booster to land in 3 out of 18 samples in the test set

CONCLUSION

- Using the models from this report SpaceY can predict when SpaceX will successfully land the 1st stage booster with 83.3% accuracy
- SpaceX public statements indicate the 1st stage booster costs upwards of $15 million to build
- This will enable SpaceY to make more informed bids against SpaceX, since they will have a good idea when to expect the SpaceX bid to include the cost of a sacrificed 1st stage booster
- With a list price of $62 million per launch, sacrificing the $15+ million 1st stage, would put the SpaceX bid at upwards of $77 million
- Biggest opportunities going forward to make even more informed bids:
  - Freeze the best performing combination of model and hyperparameters and re-fit using the whole dataset instead of just the training data
    - Potentially better than using only part of the data to fit the model, but you would no longer be able to measure the accuracy of the resulting model
  - Incorporate additional launch data to the dataset and model as it becomes available
  - Subdivide the current model into two models
    - Predict if SpaceX will ATTEMPT to land the 1st stage
    - Predict if SpaceX will SUCCEED in their attempt
  - Create a related model that predicts if SpaceX will launch using a previously-flown 1st stage booster
    - Would enable SpaceY to take into account when the SpaceX bid would likely include a discount

APPENDIX

- Notebooks to recreate dataset, analysis, and models:
  - https://github.com/brt-h/Applied-Data-Science-Capstone/blob/main/Hands-on%20Lab_%20Complete%20the%20Data%20Collection%20API%20Lab.ipynb
  - https://github.com/brt-h/Applied-Data-Science-Capstone/blob/main/Hands-on%20Lab_%20Data%20Collection%20with%20Web%20Scraping.ipynb
  - https://github.com/brt-h/Applied-Data-Science-Capstone/blob/main/Hands-On%20Lab_%20Data%20Wrangling.ipynb
  - https://github.com/brt-h/Applied-Data-Science-Capstone/blob/main/Hands-on%20Lab%20Complete%20the%20EDA%20with%20SQL.ipynb
  - https://github.com/brt-h/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb
  - https://github.com/brt-h/Applied-Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.ipynb
  - https://github.com/brt-h/Applied-Data-Science-Capstone/blob/main/spacex_dash_app.py
  - https://github.com/brt-h/Applied-Data-Science-Capstone/blob/main/Hands-on%20Lab_%20Build%20an%20Interactive%20Dashboard%20with%20Ploty%20Dash.ipynb
  - https://github.com/brt-h/Applied-Data-Science-Capstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb
- Acknowledgments
  - Thank you to Joseph Santarcangelo at IBM for creating the course and materials
  - Thank you to Lakshmi Holla at IBM for assisting me with questions and troubleshooting
- References
  - https://aviationweek.com/defense-space/space/podcast-interview-spacexs-elon-musk
    - Interview with Elon Musk where he discloses the 1st stage booster to cost upwards of $15 million
  - https://datascience.stackexchange.com/a/33050
    - Explanation of why you would rebuild your model using the full dataset
  - https://www.spacex.com/vehicles/falcon-9/
    - Source of SpaceX's advertised $62 million launch price

Thank you!