



IBM Developer  
SKILLS NETWORK

# Winning Space Race with Data Science

Daniel Cowan  
August 19, 2024



# Outline

---

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

# Executive Summary

---

- To better predict launch cost, we aim to predict whether the first stage of a launch will land successfully after launch
- After fetching data, we use exploratory data analysis methods to review possible correlations in the data and explore trends
- To make our predictions, we analyze the accuracy of four models: logistic regression, support vector machine, decision tree classifier, and k-nearest neighbors
- Our analysis finds that all models perform similarly on this task, achieving about 83% accuracy in determining whether the first stage will land successfully

# Introduction

---

- Launch costs are heavily driven by whether the first stage of the launch can be recovered
- If we can predict whether the first stage is likely to land successfully and therefore be recovered, we can more accurately estimate the launch cost
- In this project, we aim to use SpaceX historical launch data to predict whether the first stage will land successfully
  - We can use the predictions further down the line to predict launch costs



Section 1

# Methodology

# Methodology

---

- Data collection methodology:
  - Data was collected from APIs and a Wikipedia table
- Perform data wrangling
  - Categorical data was one-hot encoded, a binary target variable was created, null data replaced, and non-Falcon9 launches were filtered out
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logistic regression, SVM, decision tree, and KNN tested

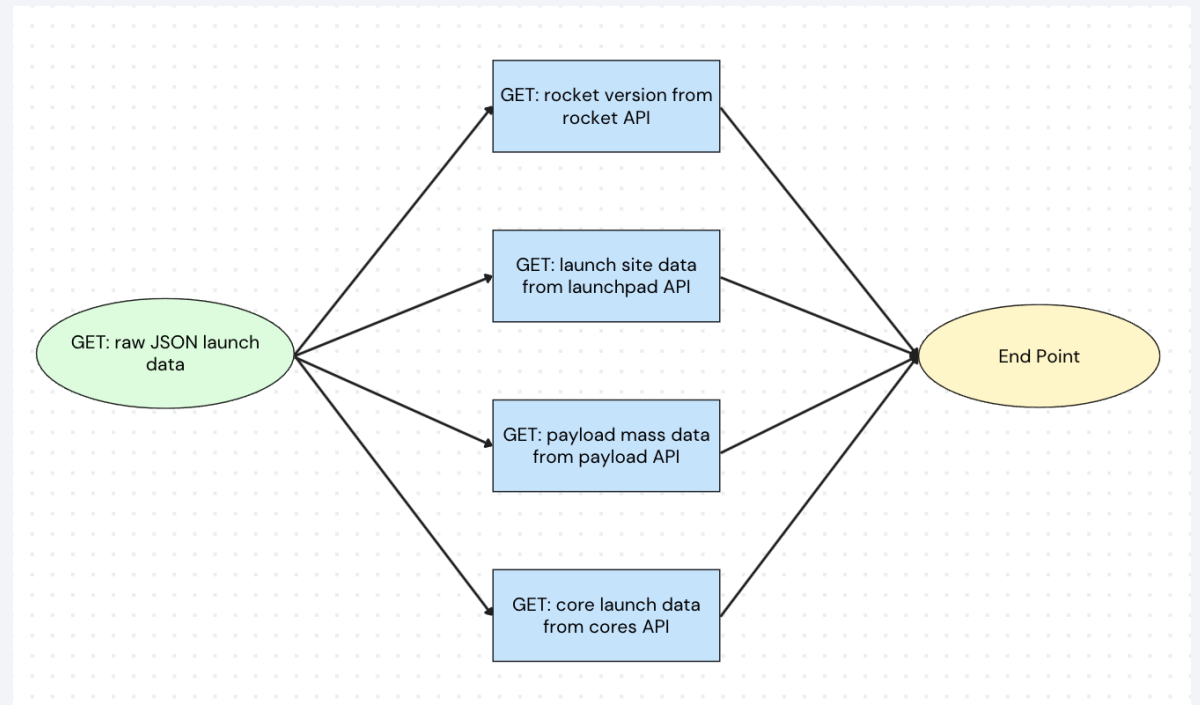
# Data Collection

---

- Data was collected from two main sources:
  - SpaceX APIs (calls using the requests library)
    - Retrieved rocket version, launch site, payload mass, and core launch data from various APIs
  - Wikipedia (scraping using BeautifulSoup)
    - Launch data, including launch and landing outcomes, scraped from Falcon9 Launches Wikipedia table

# Data Collection – SpaceX API

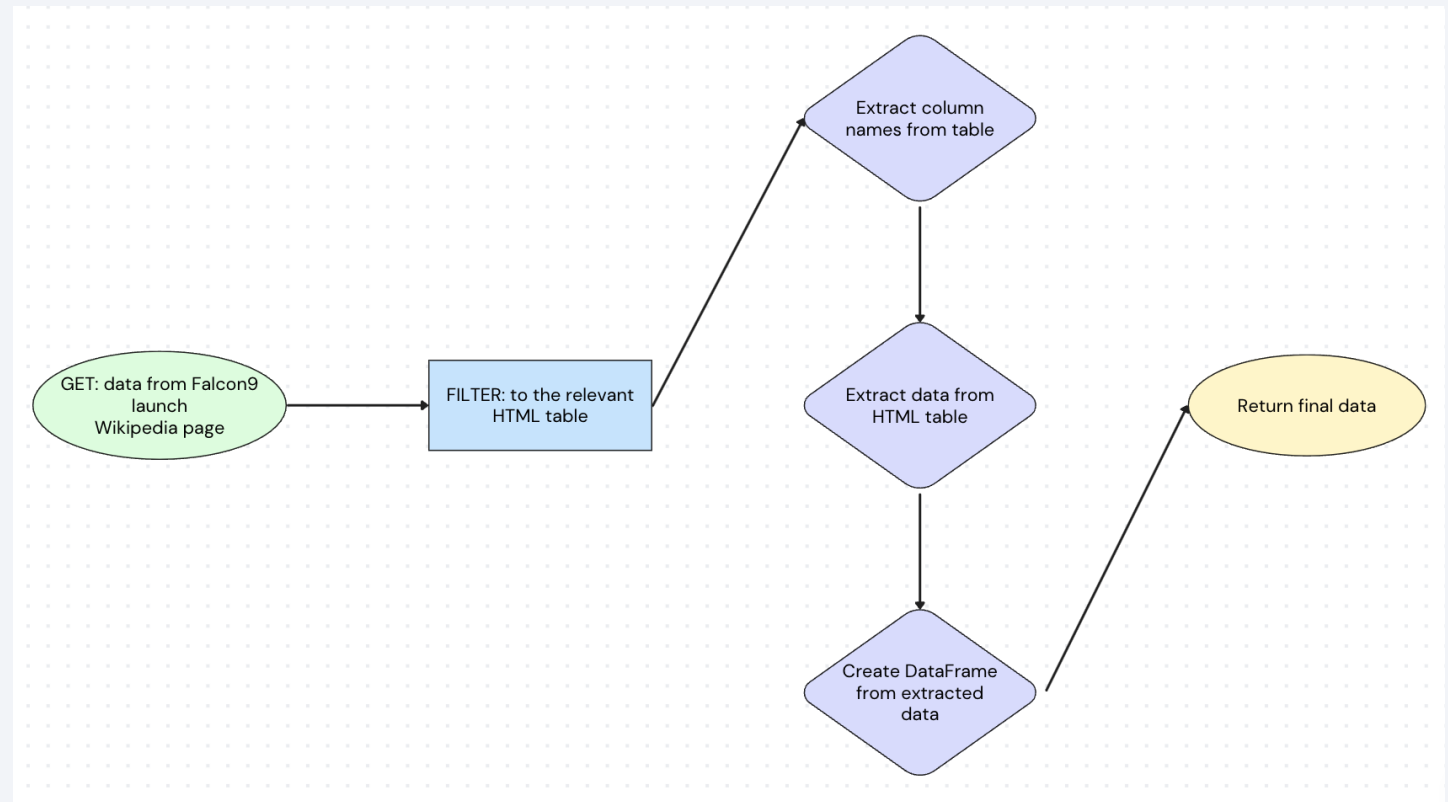
- As shown, data was collected from several APIs, and was prepped for visualization and modeling using data wrangling methods
- <https://github.com/dcstats/DSP-Cert/blob/main/Capstone/jupyter-labs-spacex-data-collection-api.ipynb>





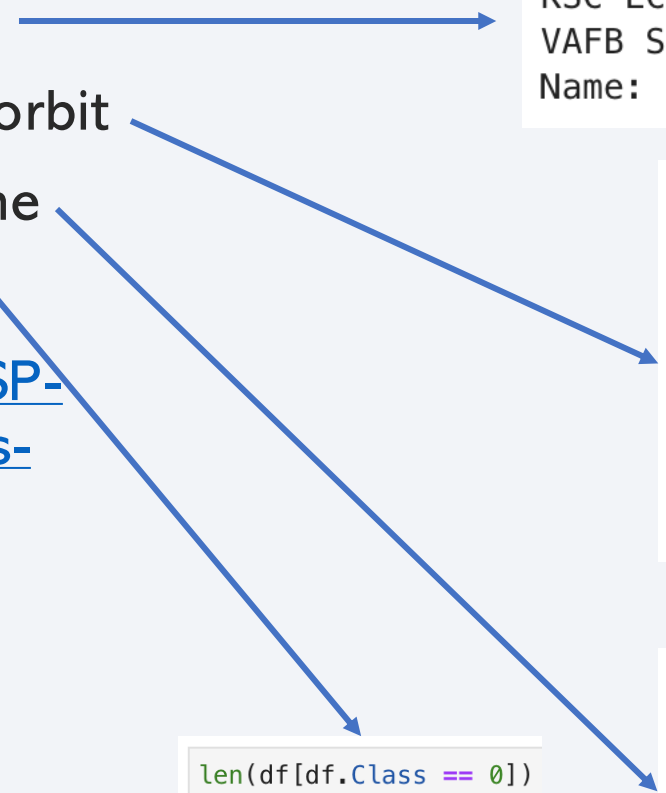
# Data Collection - Scraping

- Data was collected from the Falcon9 Launch Wikipedia page, and was prepped for visualization and modeling using BeautifulSoup and data wrangling methods
- <https://github.com/dcstats/DSP-Cert/blob/main/Capstone/jupyter-labs-webscraping.ipynb>



# Data Wrangling

- Number of launches per site
- Number of launches for each orbit
- Totals of each landing outcome
- Create landing outcome label
- <https://github.com/dcstats/DSP-Cert/blob/main/Capstone/labs-jupyter-spacex-Data%20wrangling.ipynb>



```
CCAFS SLC 40      55
KSC LC 39A        22
VAFB SLC 4E       13
Name: LaunchSite, dtype: int64
```

```
GTO      27
ISS      21
VLEO     14
PO        9
LEO        7
SSO        5
MEO        3
ES-L1      1
HEO        1
SO         1
GEO        1
Name: Orbit, dtype: int64
```

```
len(df[df.Class == 0])
30
```

```
True ASDS      41
None None      19
True RTLS      14
False ASDS      6
True Ocean      5
False Ocean     2
None ASDS       2
False RTLS      1
Name: Outcome, dtype: int64
```

# EDA with Data Visualization

---

- Visualizations created:
  - Scatterplot: Flight number vs. launch site to review whether launch locations changed as time passed
  - Scatterplot: Payload mass vs launch site to review whether larger payloads were launched at certain locations
  - Bar chart: orbit vs. landing success rate to review whether launches into certain orbits had more landing success
  - Scatterplot: Flight number vs orbit to review whether types of orbits changed over time
  - Scatterplot: Payload mass vs orbit to review whether orbit was affected by mass of load
  - Line plot: Date vs landing success rate to review how success rate changed over time
- Feature engineering: dummy variables for categorical features
- <https://github.com/dcstats/DSP-Cert/blob/main/Capstone/edadataviz.ipynb>

# EDA with SQL

---

- EDA tasks include viewing:
  - 1) Unique launch sites, 2) 5 records with launch sites starting with the string 'CCA', 3) total payload mass carried by customer NASA (CRS), 4) average payload mass carried by booster version 'F9 v1.1', 5) the date when the first successful ground pad landing was completed, 6) the names of the boosters carrying payload mass between 4,000 and 6,000 kg with successful drone ship landings, 7) the counts of each landing outcome, 8) the names of booster versions which have carried the maximum payload mass, 9) selected features for records in 2015 with failed drone ship landings, and 10) the ranked count of landing outcomes for records between June 4, 2010 and March 20, 2017
- [https://github.com/dcstats/DSP-Cert/blob/main/Capstone/jupyter-labs-eda-sql-coursera\\_sqlite.ipynb](https://github.com/dcstats/DSP-Cert/blob/main/Capstone/jupyter-labs-eda-sql-coursera_sqlite.ipynb)

# Build an Interactive Map with Folium

---

- For exploration purposes, including examining proximity of the launch sites to coasts, the following were visualized with Folium:
  - All launch sites marked with Markers
  - Landing success outcomes for each launch site with MarkerClusters
  - Distances from launch site to coast
  - Distances from launch site to highway
- [https://github.com/dcstats/DSP-Cert/blob/main/Capstone/lab\\_jupyter\\_launch\\_site\\_location.ipynb](https://github.com/dcstats/DSP-Cert/blob/main/Capstone/lab_jupyter_launch_site_location.ipynb)



# Build a Dashboard with Plotly Dash

---

- Pie Chart:
  - Proportion of successful landings by launch site
  - Proportion of landing outcomes (successful/unsuccessful) for a selected launch site
- Scatter Plot
  - Payload mass vs landing outcome to examine whether the payload size of the launch affects whether the first stage is able to land successfully
- [https://github.com/dcstats/DSP-Cert/blob/main/Capstone/spacex\\_dash\\_app.py](https://github.com/dcstats/DSP-Cert/blob/main/Capstone/spacex_dash_app.py)

# Predictive Analysis (Classification)

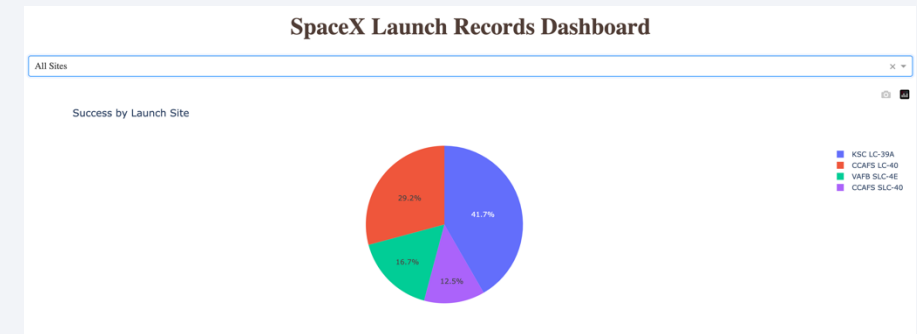
---

- Models built and evaluated:
  - Logistic regression
  - Support Vector Machine (SVM)
  - Decision Tree Classifier
  - K-Nearest Neighbors (KNN)
- Models were trained on a training set and tested for accuracy score on a separate test set of data
- [https://github.com/dcstats/DSP-Cert/blob/main/Capstone/SpaceX\\_Machine%20Learning%20Prediction\\_Part\\_5.ipynb](https://github.com/dcstats/DSP-Cert/blob/main/Capstone/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb)

# Results

---

- EDA
  - Orbit type seems to have some effect on landing success
  - Landing success increased over time, likely due to improving infrastructure, which is not captured in the data
  - In data dating back to 2010, a ground pad landing was not achieved until 2015
- Interactive analytics
- Predictive analysis
  - The models all achieved the same test set accuracy, indicating:
    - The type of model is less important for achieving predictive accuracy
    - There might be an accuracy ceiling for the data we have available







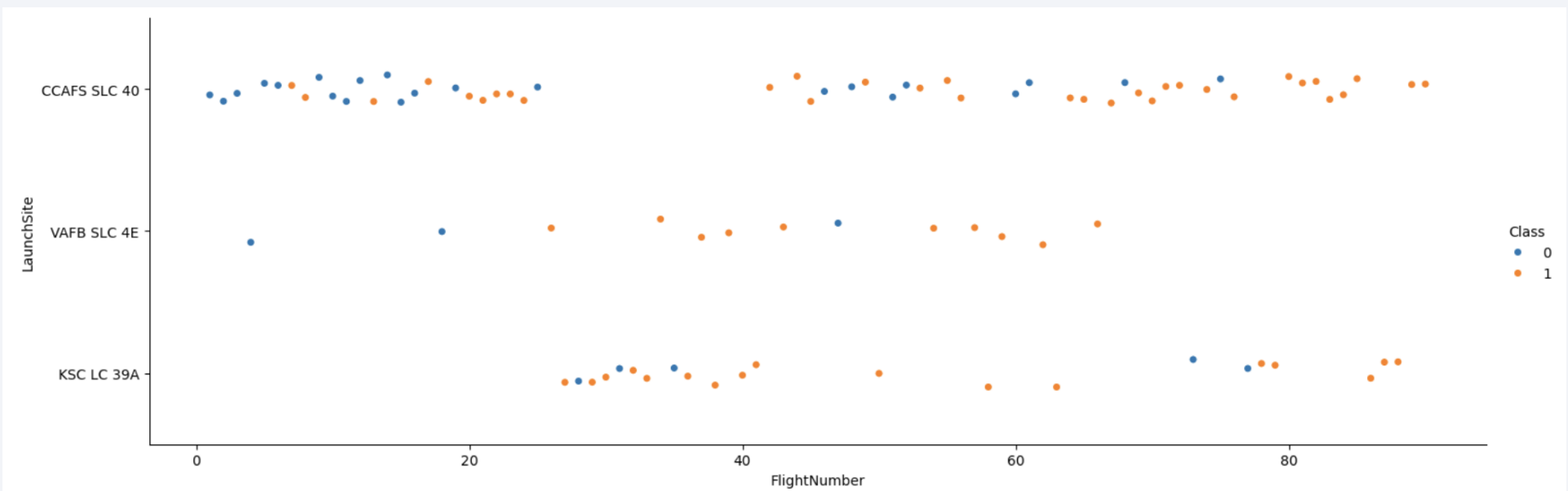
Section 2

# Insights drawn from EDA



# Flight Number vs. Launch Site

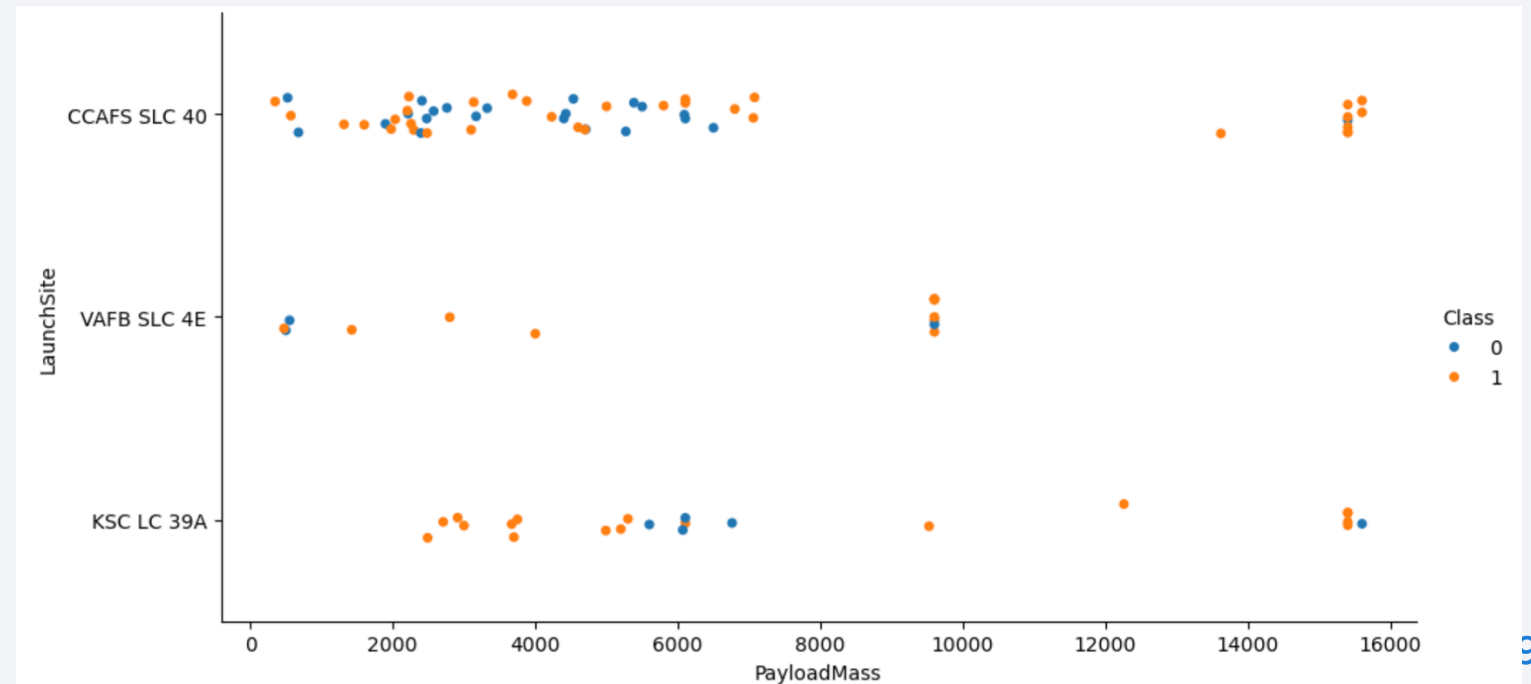
- As time passed, it's clear that landing success increased
- Early on, launches were held at the CCAFS SLC 40 launch site, and landings there were generally unsuccessful





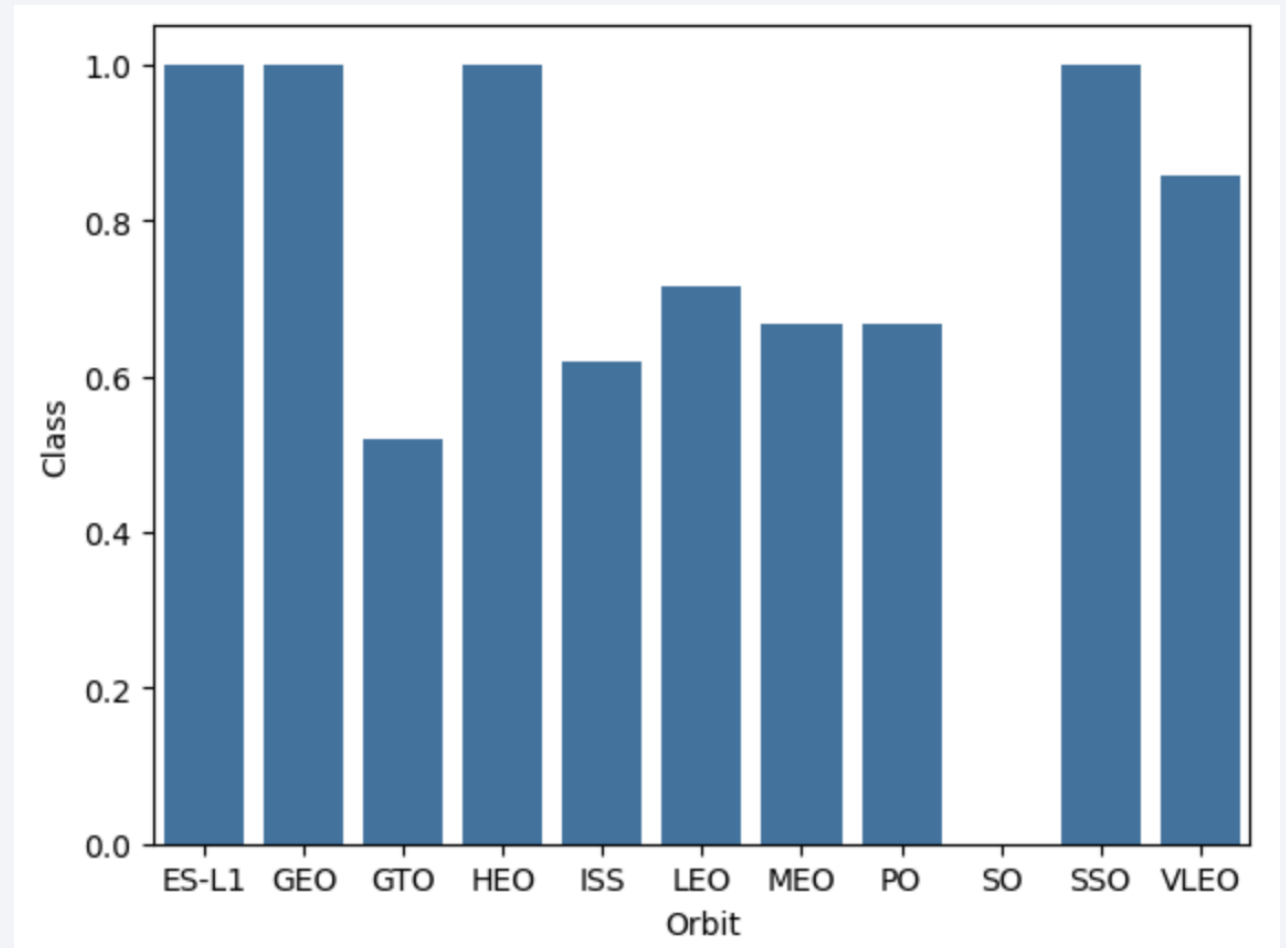
# Payload vs. Launch Site

- Most launches have smaller payload masses
- In general, as payload size increases, so does landing success rate
- This could be a bias,  
as there is more incentive  
to protect large payloads



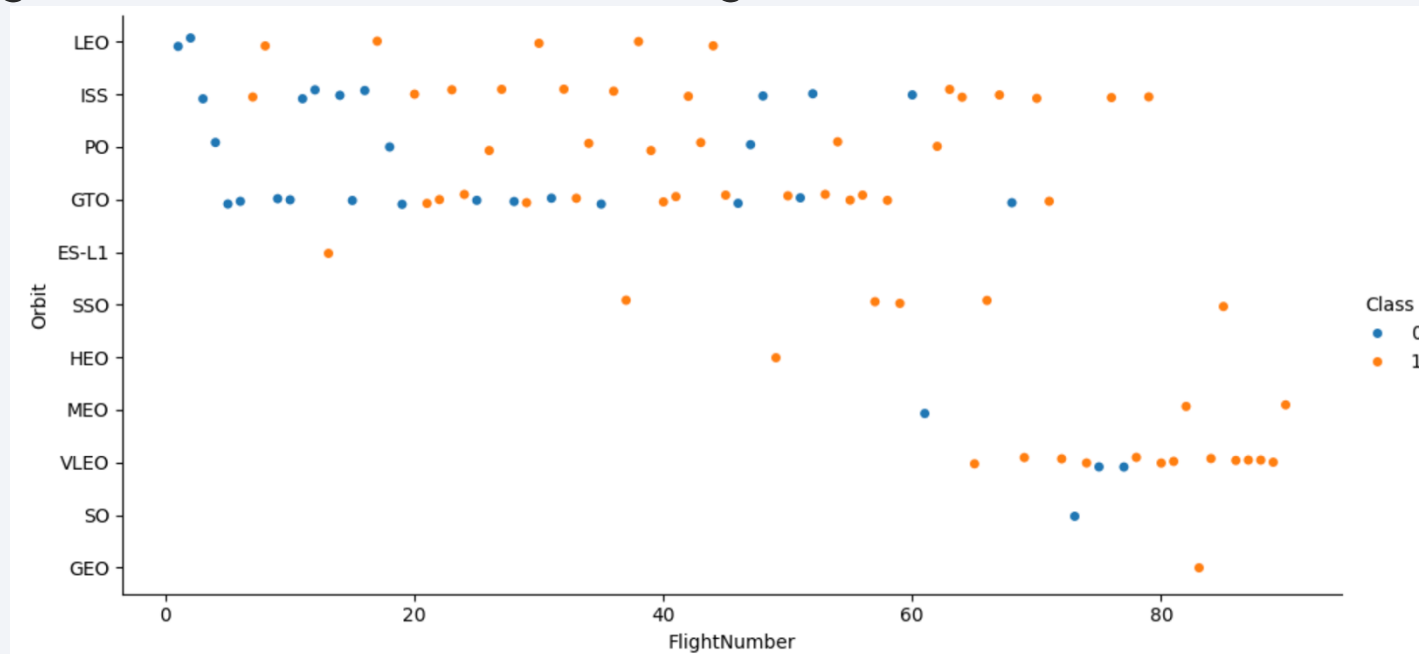
# Success Rate vs. Orbit Type

- SO orbits have yet to see a successful landing
- ES-L1, GEO, HEO, and SSO orbits all have perfect landing success rates



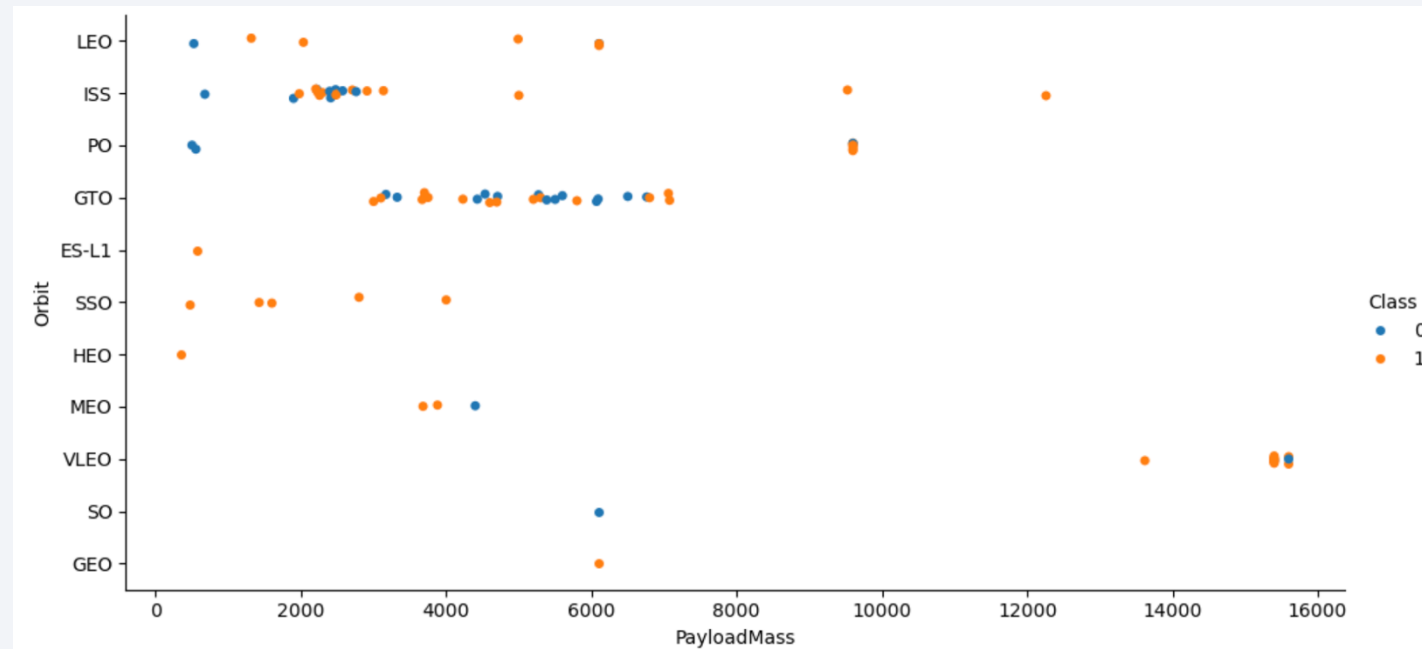
# Flight Number vs. Orbit Type

- As time increased, SpaceX switched to what are likely more complicated orbits; from LEO/ISS to VLEO/SSO
- Most early flights had unsuccessful landings, but success increased with time



# Payload vs. Orbit Type

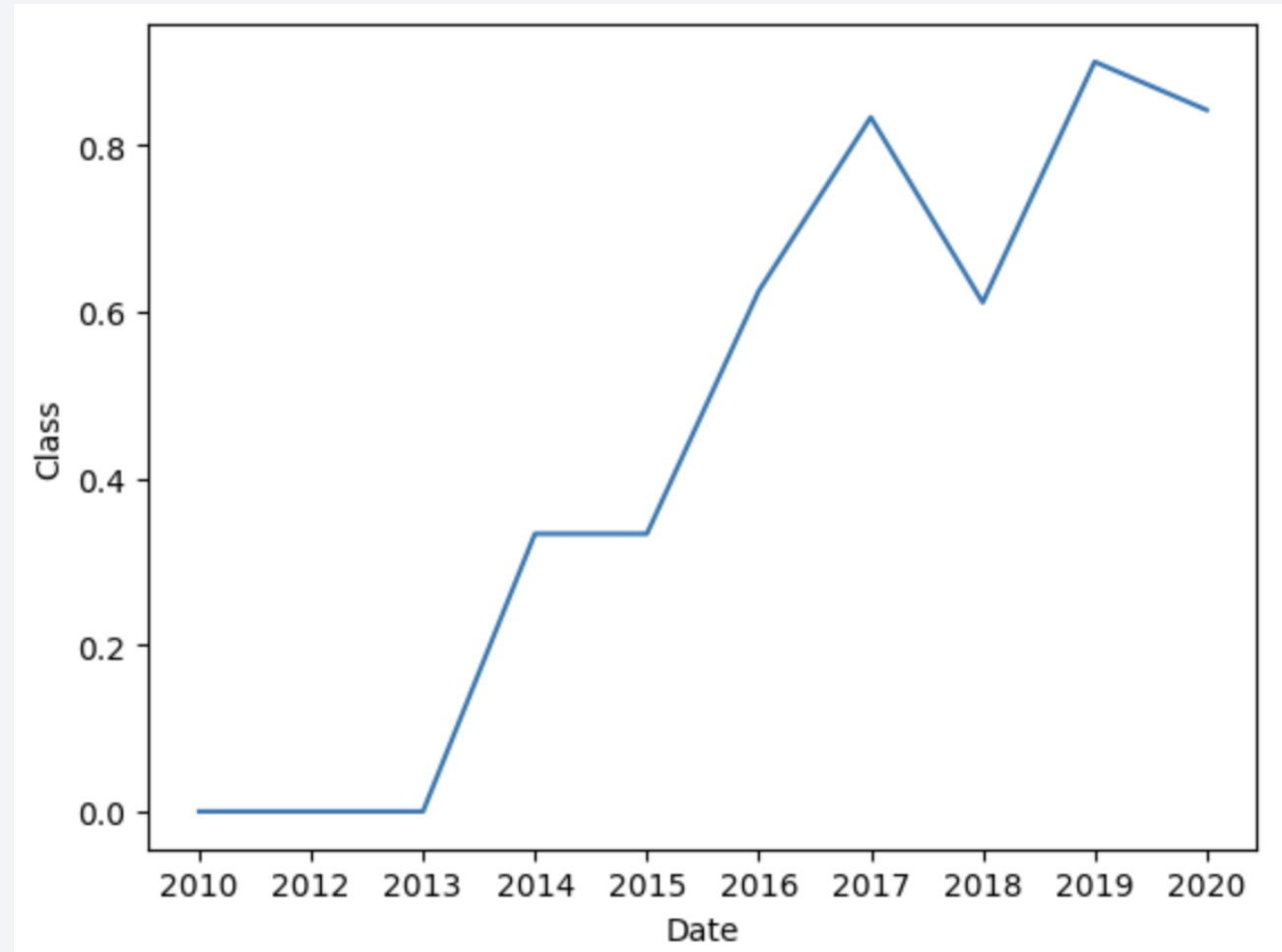
- Most launches have lower payload masses
- VLEO orbits have the highest payloads; LEO and SSOs have among the lowest



# Landing Success Yearly Trend

---

- Landing success clearly has increased over time
- Innovations in landing technology clearly has had a positive effect on being able to land the first stage
- Current success rates hover around 80%





# All Launch Site Names

---

- Of the ~100 records, there are 4 unique launch sites

```
[11]: %sql select distinct Launch_Site from SPACEXTABLE
      * sqlite:///my_data1.db
Done.
[11]: Launch_Site
      CCAFS LC-40
      VAFB SLC-4E
      KSC LC-39A
      CCAFS SLC-40
```

# Launch Site Names Begin with 'CCA'

- All 5 of the shown records are from launch site CCAFS LC-40
- None of the shown launches ended in successful landing

```
[14]: %sql select * from SPACEXTABLE where Launch_Site like "CCA%" limit 5
```

```
* sqlite:///my_data1.db  
Done.
```

```
[14]:
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

# Total Payload Mass

---

- NASA (CRS) carried a total payload mass of exactly 45,596 kg
- Note that CRS is not the only program run by NASA that employed Space X's services

```
[18]: %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = "NASA (CRS)"
      * sqlite:///my_data1.db
      Done.
[18]: sum(PAYLOAD_MASS__KG_)
      45596
```

# Average Payload Mass by F9 v1.1

---

- Booster version F9 v1.1 carries an average payload of 2,928.4 kg

```
[21]: %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version = "F9 v1.1"
      * sqlite:///my_data1.db
      Done.
[21]: avg(PAYLOAD_MASS__KG_)
      2928.4
```

# First Successful Ground Landing Date

---

- The first successful ground pad landing was on December 22, 2015

```
[25]: %sql select min(Date) from SPACEXTABLE where Landing_Outcome = "Success (ground pad)"
      * sqlite:///my_data1.db
      Done.
[25]: min(Date)
      2015-12-22
```



## Successful Drone Ship Landing with Payload between 4000 and 6000

---

- There are 4 boosters with payloads between 4,000 and 6,000 kg which had successful drone ship landings

```
[19]: %sql select distinct Booster_Version from SPACEXTABLE \
      where Landing_Outcome = "Success (drone ship)" and PAYLOAD_MASS_KG_ between 4000 and 6000
      * sqlite:///my_data1.db
      Done.
```

```
[19]: Booster_Version
```

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

# Total Number of Successful and Failure Mission Outcomes

---

- All but 1 mission was listed as successful
- Note that this is different from the landing outcome; far more landings were unsuccessful than missions

```
[17]: %sql select Mission_Outcome, count(Mission_Outcome) from SPACEXTABLE group by Mission_Outcome
```

```
* sqlite:///my_data1.db  
Done.
```

```
[17]:
```

Mission_Outcome	count(Mission_Outcome)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

# Boosters Carried Maximum Payload

---

- A total of 12 unique boosters were able to carry the maximum payload
- Note that only F9 B5 boosters were able to meet this criteria

```
[32]: %sql select distinct Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)
* sqlite:///my_data1.db
Done.
```

[32]:	Booster_Version
	F9 B5 B1048.4
	F9 B5 B1049.4
	F9 B5 B1051.3
	F9 B5 B1056.4
	F9 B5 B1048.5
	F9 B5 B1051.4
	F9 B5 B1049.5
	F9 B5 B1060.2
	F9 B5 B1058.3
	F9 B5 B1051.6
	F9 B5 B1060.3
	F9 B5 B1049.7

# 2015 Launch Records

---

- There were just 2 failed drone ship landings in 2015; one in January and one in April
- Both launched from the same site but used different boosters

```
[44]: %sql select substr(Date, 6,2) as Month, Landing_Outcome, Booster_Version, Launch_Site \
      from SPACEXTABLE \
      where Landing_Outcome = "Failure (drone ship)" \
      and substr(Date,0,5) = '2015'
```

```
* sqlite:///my_data1.db
Done.
```

```
[44]:
```

	Month	Landing_Outcome	Booster_Version	Launch_Site
	01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
	04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- Most landings were not attempted at all; this is likely in the early years before technology to attempt landings was as advanced as it is now
- Drone ship landing attempts followed in second

```
[40]: %sql select Landing_Outcome, count(Landing_Outcome) as ct \
      from SPACEXTABLE \
      where Date between "2010-06-04" and "2017-03-20" \
      group by Landing_Outcome \
      order by ct desc
```

```
* sqlite:///my_data1.db
Done.
```

```
[40]:
```

Landing_Outcome	ct
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1



Section 3

# Launch Sites Proximities Analysis

# Launch Site Map

---

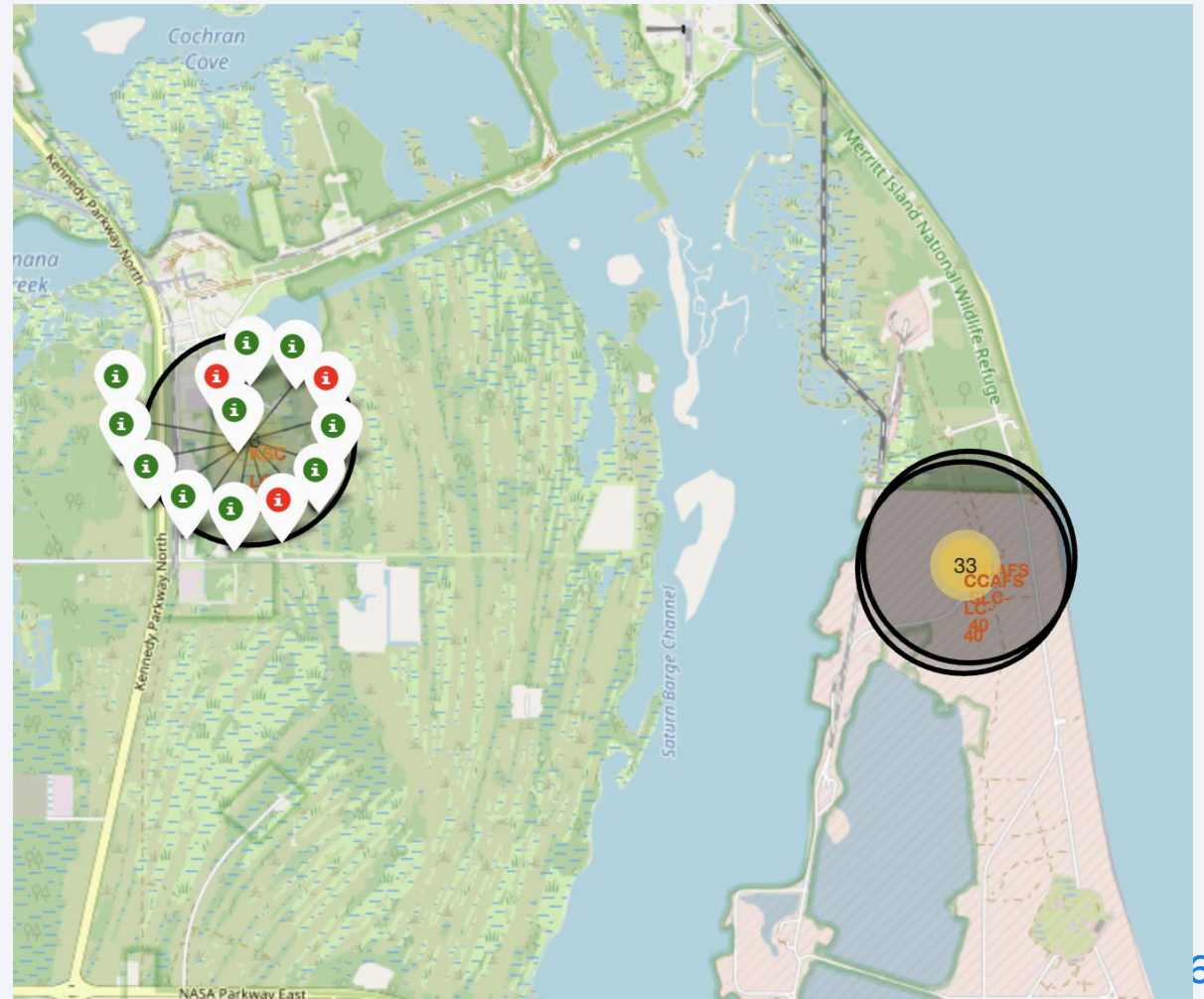
- All launch sites are exclusively coastal
- 1 site in CA, 3 in FL





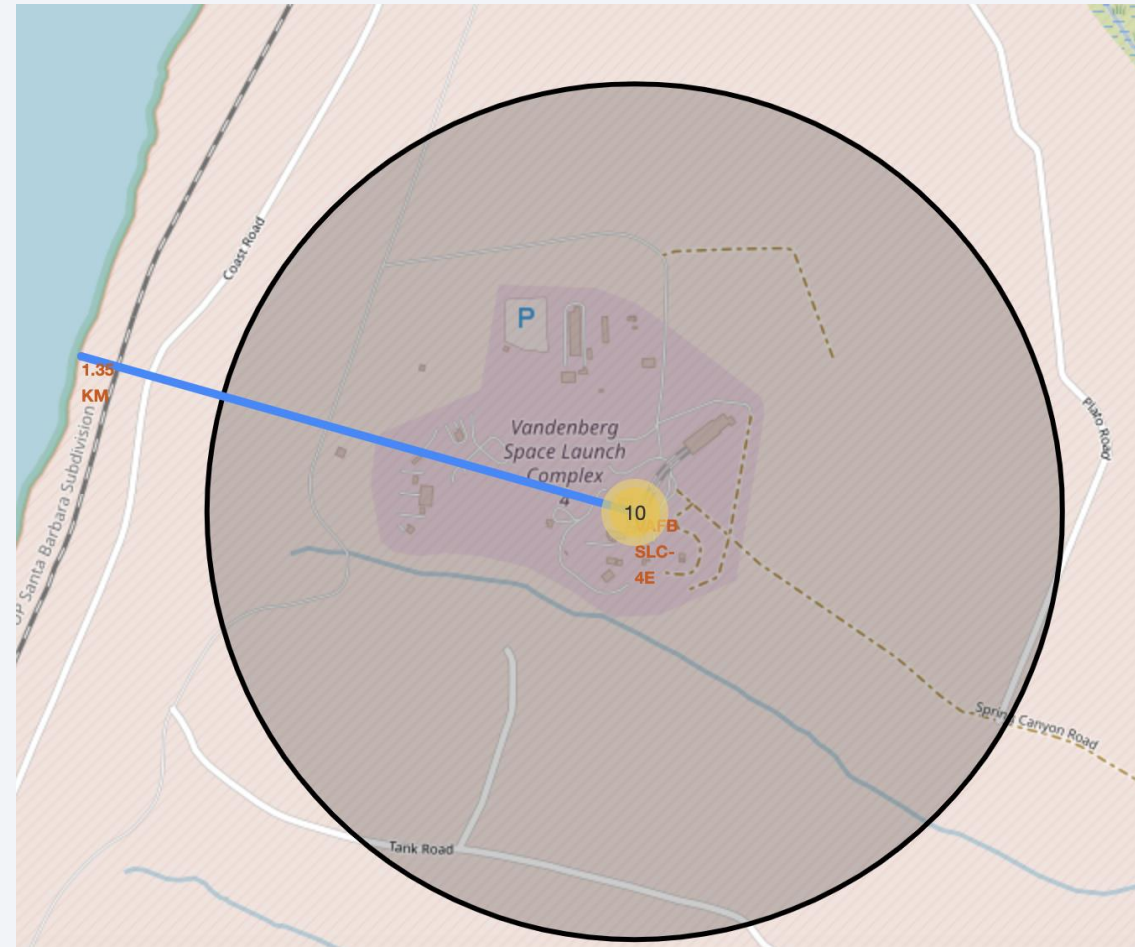
# Launch Site Success

- The more inland launch site appears to be more successful than the ones immediately on the coast
- All other sites have <50% success rates



# Launch Site Proximity to Shorelines

- As seen here, this CA launch site is approximately 1.35km from the shore of the Pacific Ocean
- This is a common theme with the four launch sites, as all 4 sites in the dataset are within 2 km of a shoreline
- This makes sense; the ocean is the best way to avoid collateral damage in the case of a failed launch





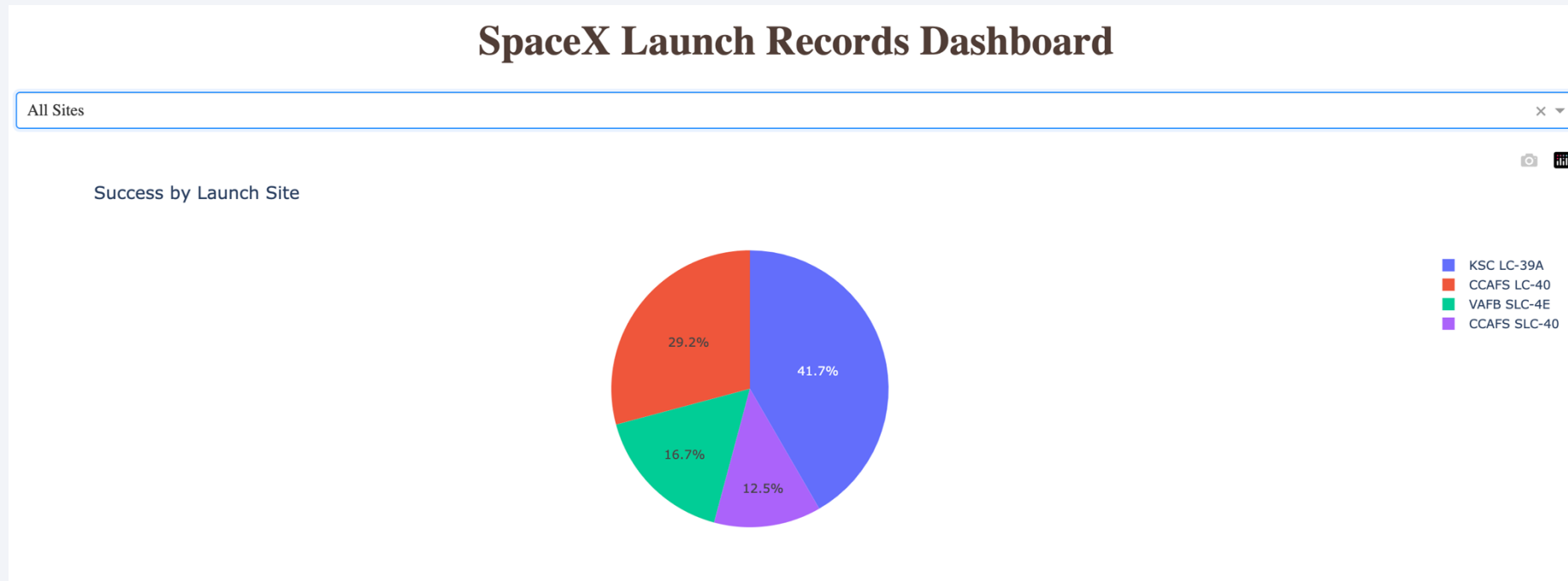


Section 4

# Build a Dashboard with Plotly Dash

# Launch Success by Launch Site

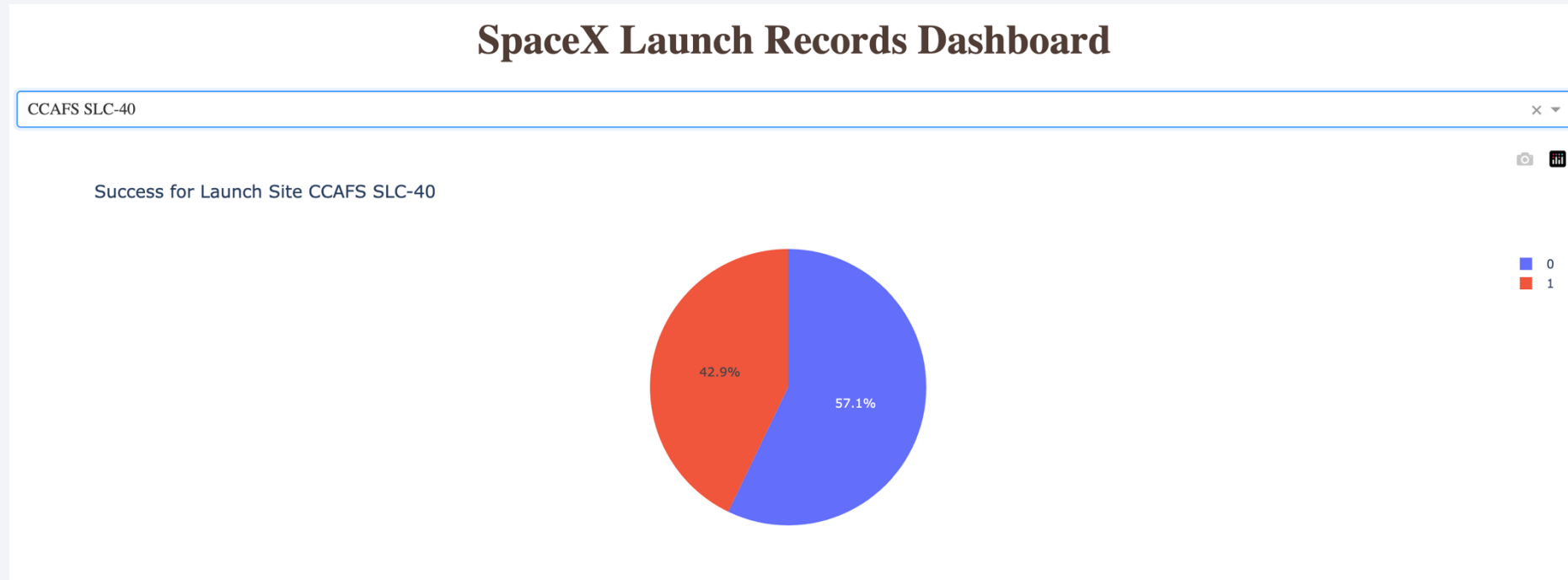
- Launch site KSC LC-39A has the most successful launches, though it should be noted that this is not the same as having the highest success rate



# Launch Site Success Percentage

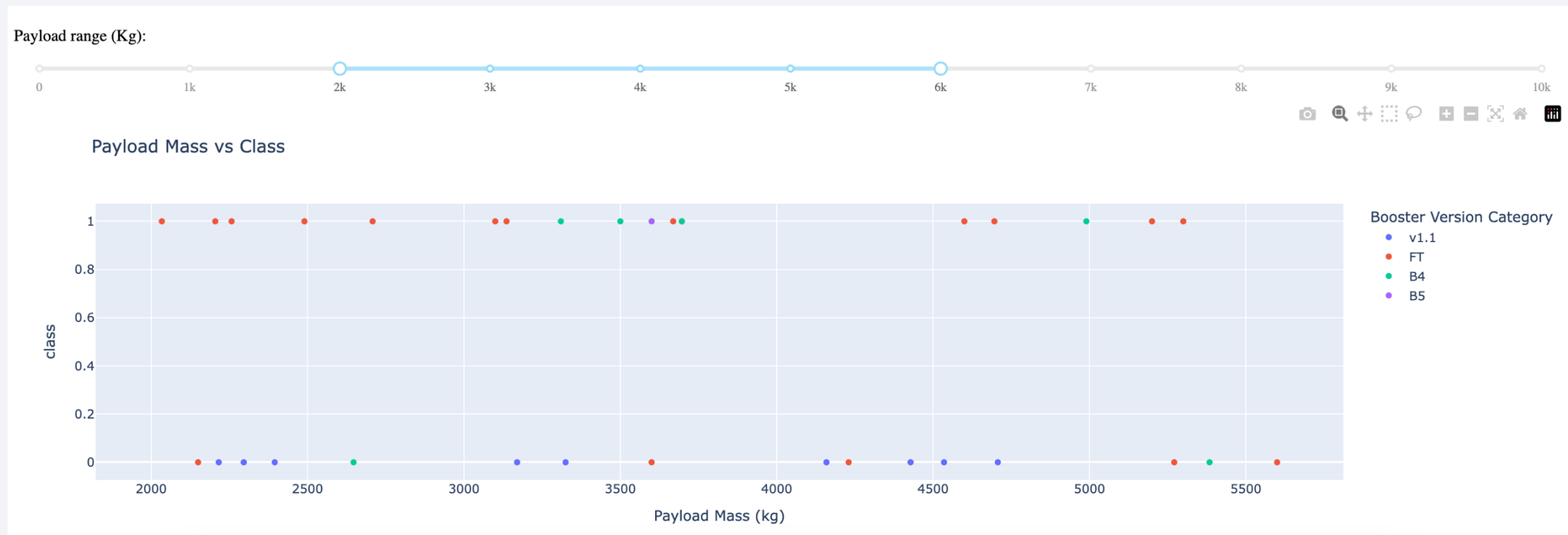
---

- Launch site CCAFS SLC-40 is the most successful, with about 43% of launches being successful



# Payload Mass vs. Landing Success

- Booster version FT seems to yield the most success
- There looks like no clear correlation between payload mass and landing success



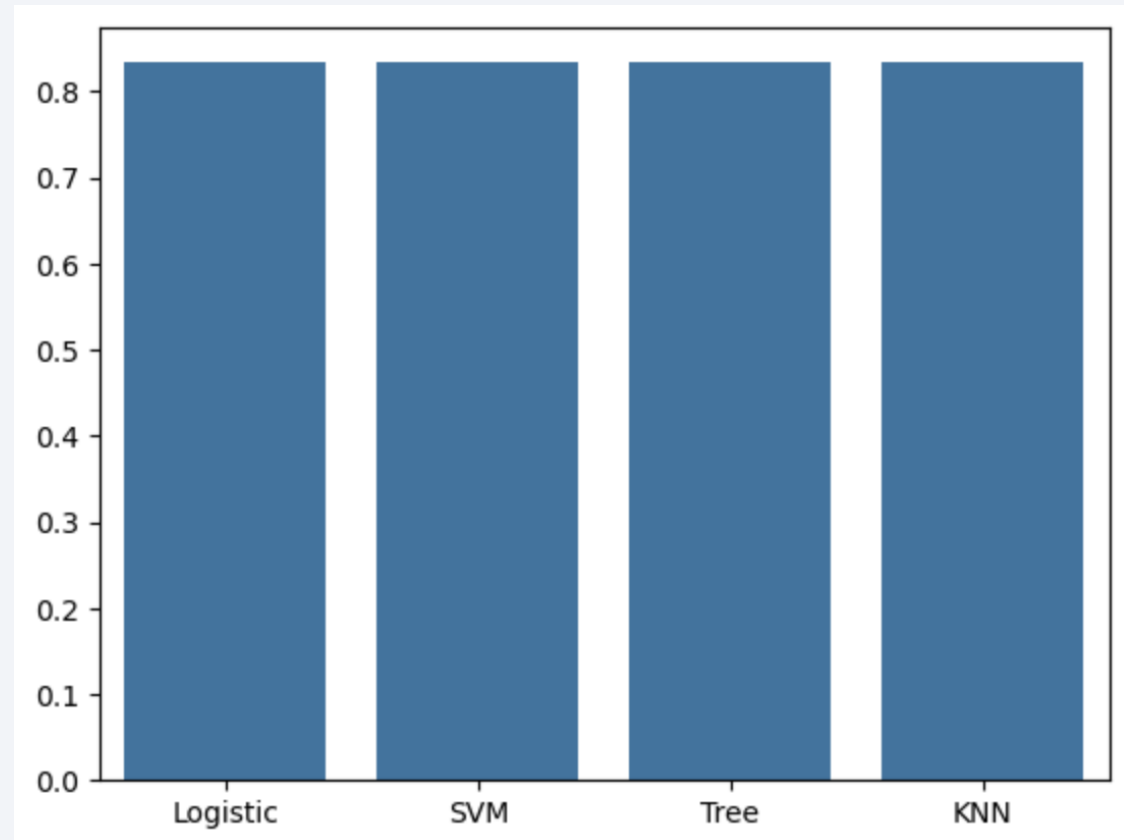
Section 5

# Predictive Analysis (Classification)

# Classification Accuracy

---

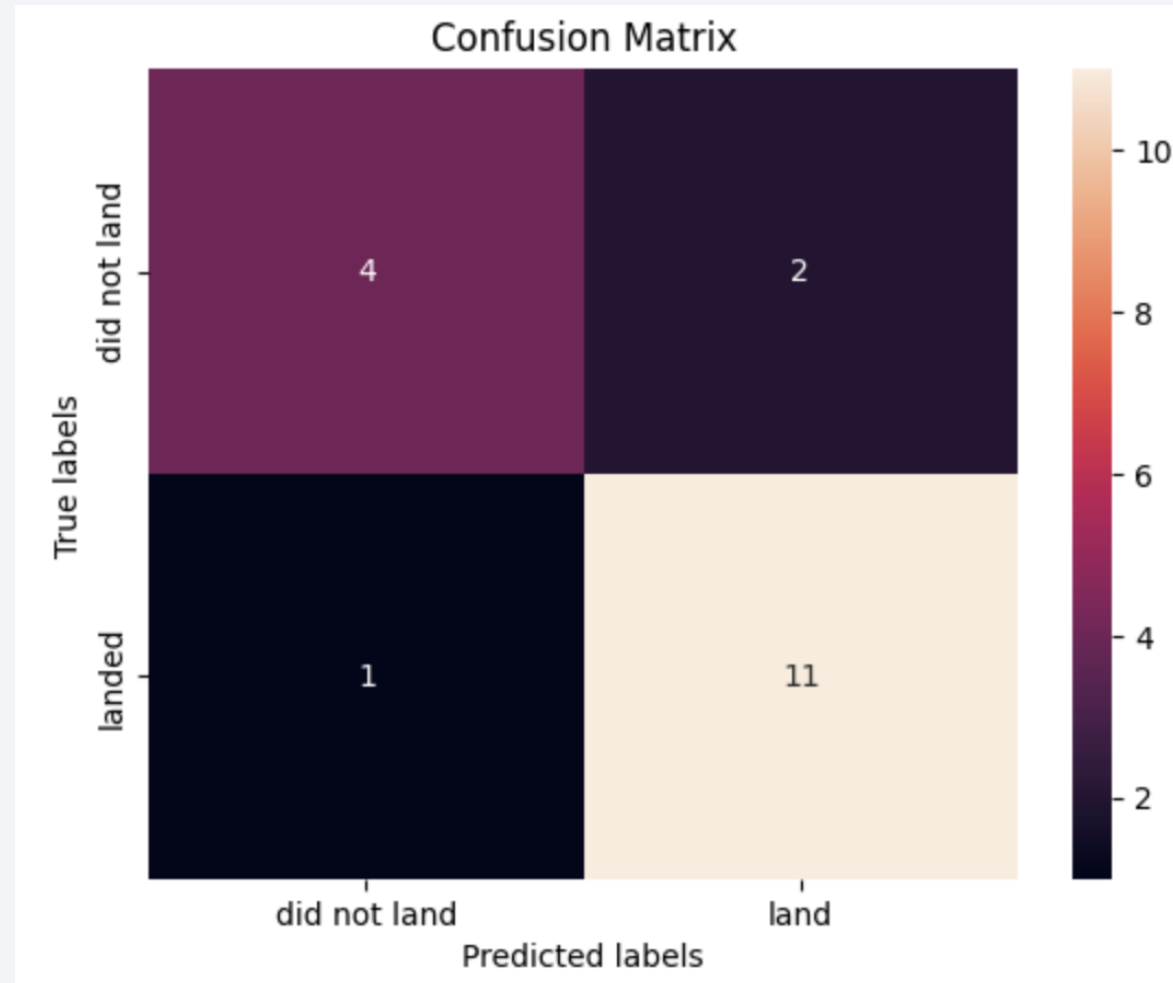
- Interestingly, all four models had the same test set accuracy of 83%
- None seem to be overfit, as average training accuracy is just higher than average test accuracy (around 84%)





# Confusion Matrix

- Confusion matrix for the Decision Tree Classifier model



# Conclusions

---

- All models performed similarly on the task
- There are likely other factors that have greater influence on landing outcome which are difficult to include in the data, such as technology development
  - Date might be a good proxy for this, as average landing success steadily increased over time
- The curse of dimensionality is something to consider, as we have around 70 features from the encoded categorical variables alone
- The accuracy is sufficient to improve cost forecasting



Thank you!

