

# Winning Space Race with Data Science

Daniil Melnik  
05-06-2024



# Структура

---

- Краткое изложение
- Введение
- Методология
- Результаты
- Заключение

# Краткое изложение

---

## Сводка методологий:

- Сбор данных через API и с веб-скрапингом
- Обработка данных
- Разведочный анализ данных с использованием SQL и визуализации данных
- Интерактивная визуальная аналитика с использованием Folium и Plotly Dash
- Прогнозирование с использованием машинного обучения

## Обобщение всех результатов:

- Полный обзор результатов, включая визуализации и точности моделей
- Ключевые выводы из анализа, выявление трендов и корреляций

# Введение

---

- **Контекст и предпосылки проекта**

Компания SpaceX рекламирует запуски ракеты Falcon 9 на своем веб-сайте по цене 62 миллиона долларов; в то время как запуски от других провайдеров стоят более 165 миллионов долларов за каждый, большая часть экономии обусловлена возможностью Space X повторно использовать первую ступень. Следовательно, если мы можем определить, успешно ли приземлится первая ступень, мы можем определить стоимость запуска. Эта информация может быть использована, если альтернативная компания захочет конкурировать с SpaceX за запуск ракеты. Цель проекта - создать конвейер машинного обучения для прогнозирования успешной посадки первой ступени.

- **Проблемы, на которые ищем ответы**

- Какие факторы определяют успешность посадки ракеты?
- Взаимодействие между различными характеристиками, определяющими успешность посадки.
- Какие рабочие условия должны быть созданы для обеспечения успешной программы посадки.

Section 1

# Methodology

# Методология

---

Методология сбора данных:

Данные от SpaceX были получены из двух источников:

- API SpaceX (<https://api.spacexdata.com/v4/rockets/>)
- Веб-скрапинг ([https://en.wikipedia.org/wiki/List\\_of\\_Falcon\\_9\\_and\\_Falcon\\_Heavy\\_launches](https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches))

Выполнена обработка данных

- Собранные данные были обогащены путем создания метки результата посадки на основе данных о результате после суммирования и анализа признаков

# Методология

---

- Проведен исследовательский анализ данных (EDA) с использованием визуализации и SQL
- Проведен интерактивный визуальный анализ с использованием Folium и Plotly Dash
- Собранные данные, были нормализованы, разделены на обучающие и тестовые наборы данных
- Проведен прогностический анализа с использованием моделей классификации, точность каждой модели оценивалась с использованием различных комбинаций параметров.

# Сбор данных

---

## SpaceX API Calls:

- Использовали API SpaceX (<https://api.spacexdata.com/v4/rockets/>) для получения данных о запусках, полезных нагрузках и информации о ракетах-носителях.

## Web Scraping:

- Применили техники веб-скрапинга для сбора дополнительных данных, недоступных через API.
- Использовали библиотеку BeautifulSoup на Python для разбора HTML-контента.
- Извлекли данные из Wikipedia ([https://en.wikipedia.org/wiki/List\\_of\\_Falcon\\_9\\_and\\_Falcon\\_Heavy\\_launches](https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches))

# Сбор данных – SpaceX API

---

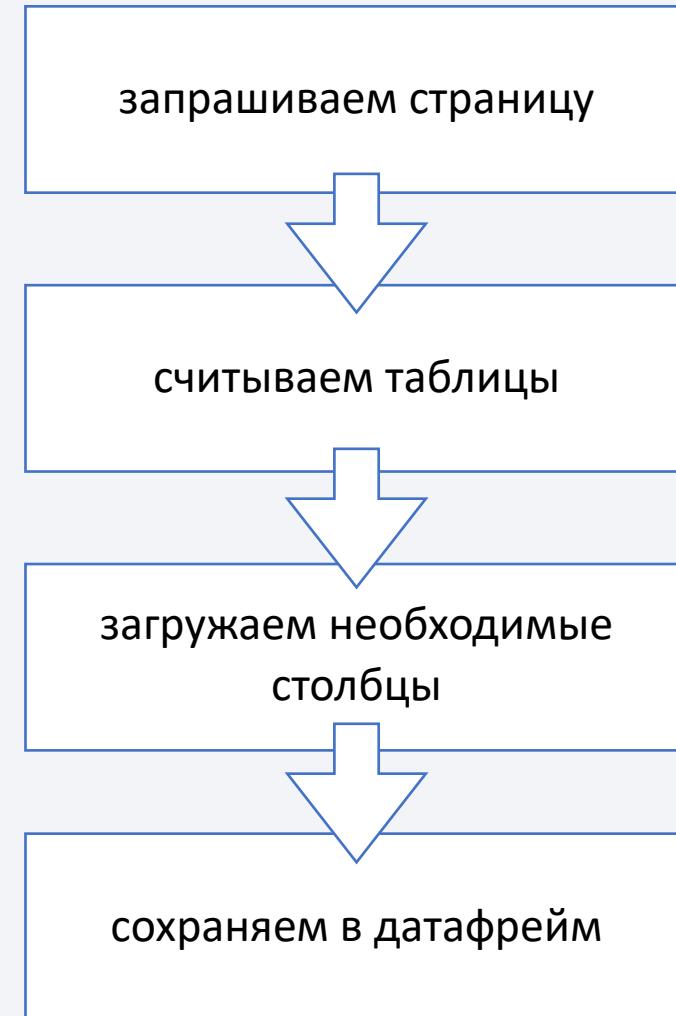
- Использовался SpaceX REST API для получения данных.
- Выполнялись запросы для получения данных о запусках, полезных нагрузках и ракетах.
- Отфильтровывались и обрабатывались данные для извлечения информации о Falcon 9



# Сбор данных - Scraping

---

- Применены техники веб-скрапинга для сбора дополнительных данных, недоступных через API.
- Использована библиотека BeautifulSoup в Python для парсинга HTML
- Использована страница на Wikipedia о запусках Falcon 9



# Обработка данных

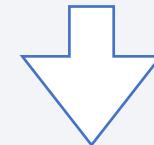
---

- Решили проблемы с отсутствующими значениями путем замены на средние значения
- Были рассчитаны сводные данные по запускам на каждой стартовой площадке, количество запусков для каждой орбиты и количество исходов миссий для каждого типа орбиты
- Был создан признак успешности посадки на основе столбца "Outcome"

проверка наличия  
пропущенных значений



замена пропущенных  
значений на среднее значение



создание признака  
успешности посадки

# EDA и визуализация данных

---

Для изучения данных использовались точечные диаграммы и столбчатые диаграммы для визуализации взаимосвязи между парами признаков:

- Масса полезной нагрузки vs. Номер полета
- Масса полезной нагрузки vs. Место запуска
- Масса полезной нагрузки vs. Номер полета
- Место запуска vs. Масса полезной нагрузки
- Орбита vs. Номер полета
- Полезная нагрузка vs. Орбита

# EDA с SQL

---

Были выполнены следующие SQL-запросы:

- Названия уникальных площадок для запуска космических миссий;
- Топ-5 площадок для запуска, названия которых начинаются со строки 'CCA';
- Общая масса полезного груза, перевезенного ракетами, запущенными NASA (CRS);
- Средняя масса полезного груза, перевезенного версией ракеты F9 v1.1;
- Дата первого успешного приземления на наземной площадке;
- Названия ракет, которые успешно приземлились на плавучей платформе и перевозили полезный груз массой от 4000 до 6000 кг;
- Общее количество успешных и неудачных исходов миссий;
- Названия версий ракет, которые перевозили максимальную массу полезного груза;
- Неудачные исходы приземления на плавучей платформе, их версии ракет и названия площадок для запуска в 2015 году; и ранг количества исходов приземления (например, неудача (плавучая платформа) или успех (наземная площадка)) между датами 2010-06-04 и 2017-03-20.

# Создание интерактивных карт с Folium

---

- На карты Folium добавили различные объекты:
- Маркеры указывают точки, такие как места запуска
- Круги показывают выделенные области вокруг определенных координат, например, Центр космических полетов имени НАСА
- Кластеры маркеров указывают на группы событий в каждой координате, например, запуски на космодроме
- Линии используются для указания расстояний между двумя координатами.

# Создание Dashboard с Plotly Dash

---

- Мы создали интерактивную панель управления с помощью Plotly Dash.
- Мы построили круговые диаграммы, показывающие общее количество запусков с определенных площадок.
- Мы построили точечный график, показывающий связь между результатом и массой полезной нагрузки (кг) для различных версий ускорителей.

# Предсказательный анализ (Классификация)

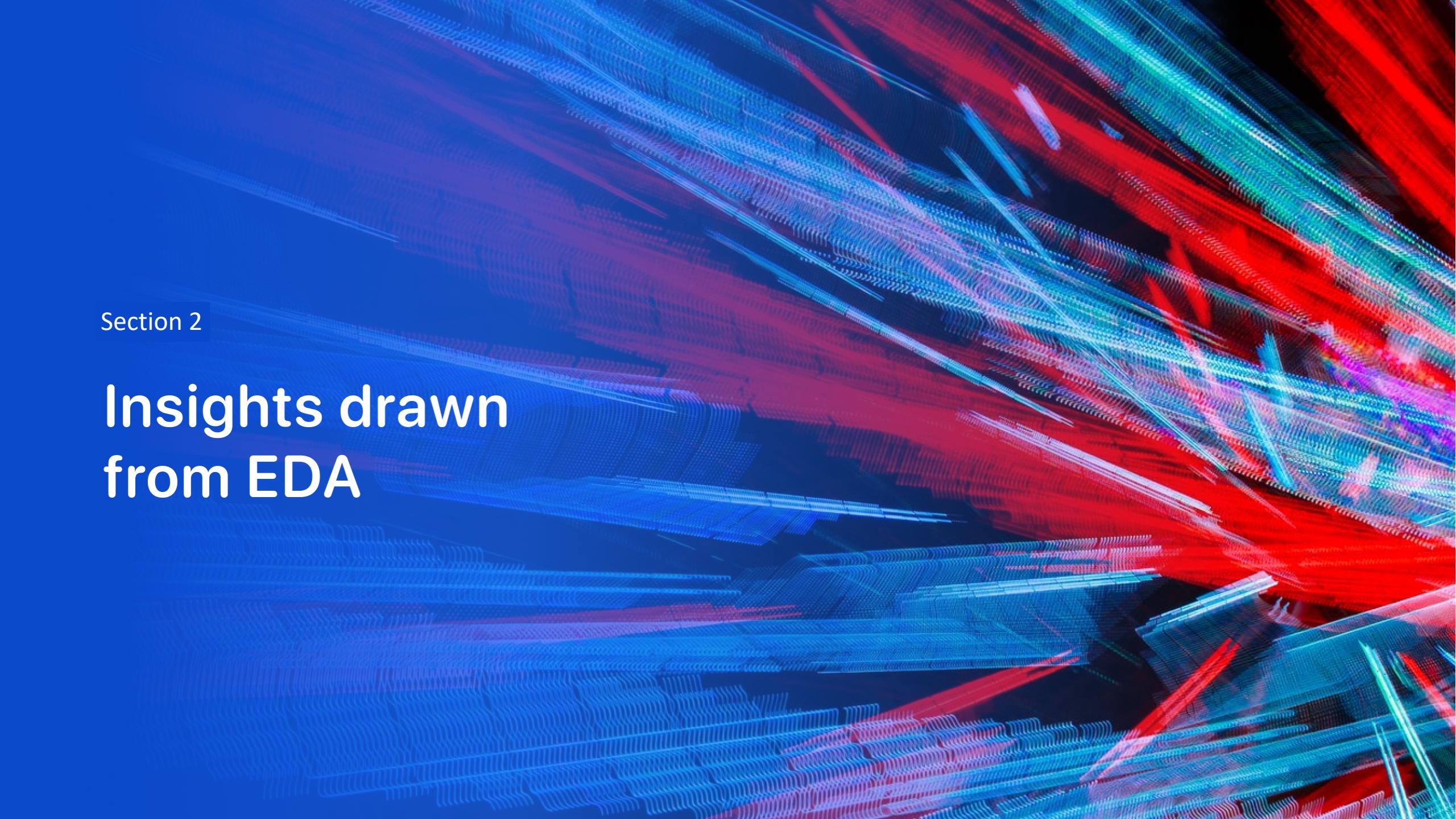
---

- Построили различные модели машинного обучения с использованием пайплайнов
- Настроили различные гиперпараметры с помощью GridSearchCV.
- Использовали accuracy как метрику для оценки моделей

# Результаты

---

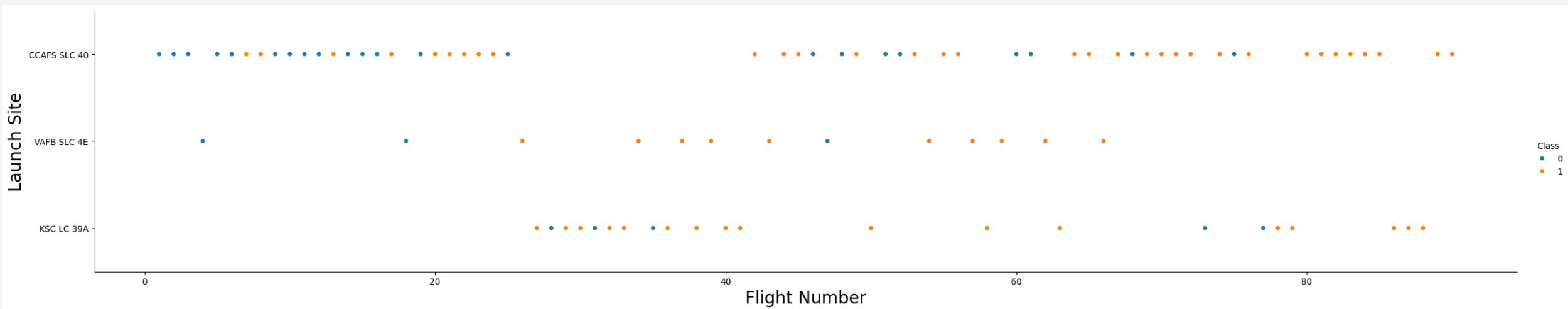
- SpaceX использует 4 разных площадки для запусков;
- Первые запуски были сделаны на площадке SpaceX и NASA;
- Средняя нагрузка для ракеты-носителя F9 v1.1 составляет 2 928 кг;
- Первый успешный посадочный исход произошел в 2015 году, через пять лет после первого запуска;
- Многие версии ракет-носителей Falcon 9 успешно посадились на плавучую платформу, имея нагрузку выше среднего;
- Две версии ракет-носителей не смогли успешно посадиться на дрон-корабли в 2015 году: F9 v1.1 B1012 и F9 v1.1 B1015;
- Количество успешных посадочных исходов становилось лучше с каждым годом.
- Используя интерактивный анализ, удалось выяснить, что места для запусков обычно находятся в безопасных местах, близко к морю, например, и имеют хорошую логистическую инфраструктуру вокруг.
- Большинство запусков происходит на площадках на восточном побережье.

The background of the slide features a complex, abstract digital visualization. It consists of numerous thin, glowing lines that create a sense of depth and motion. The lines are primarily blue and red, with some green and white highlights. They form a grid-like structure that curves and twists across the frame, resembling a 3D space or a network of data points. The overall effect is futuristic and dynamic.

Section 2

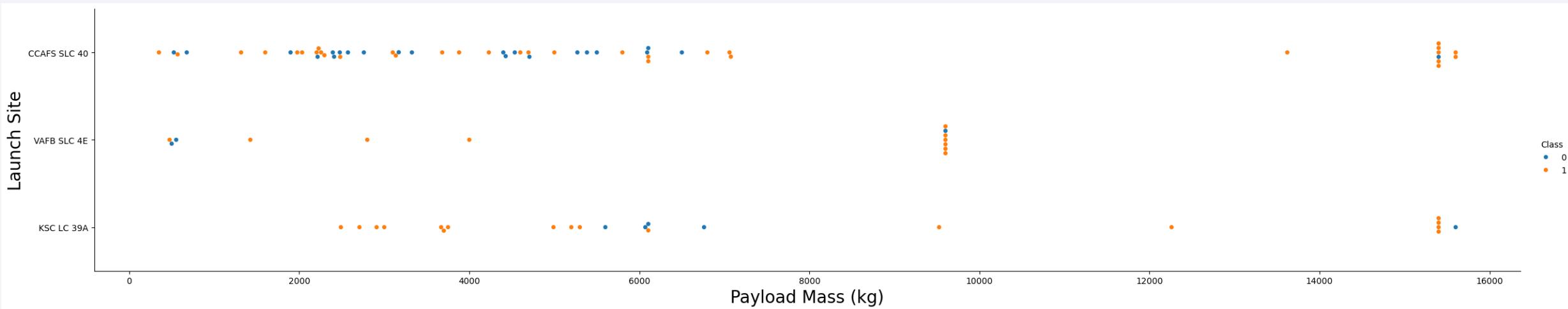
## Insights drawn from EDA

# Flight Number vs. Launch Site



Мы наблюдаем различия в процентах успешных запусков на различных площадках. В частности, на CCAFS LC-40 процент успешных запусков составляет 60%, в то время как на KSC LC-39A и VAFB SLC 4E этот процент равен 77%.

# Payload vs. Launch Site



При рассмотрении точечного графика, изображающего Грузоподъемность по отношению к Месту запуска, становится очевидным, что на площадке запуска VAFB-SLC не было запущено ракет с грузоподъемностью более 10000.

# Success Rate vs. Orbit Type

---

Наивысший уровень успеха:

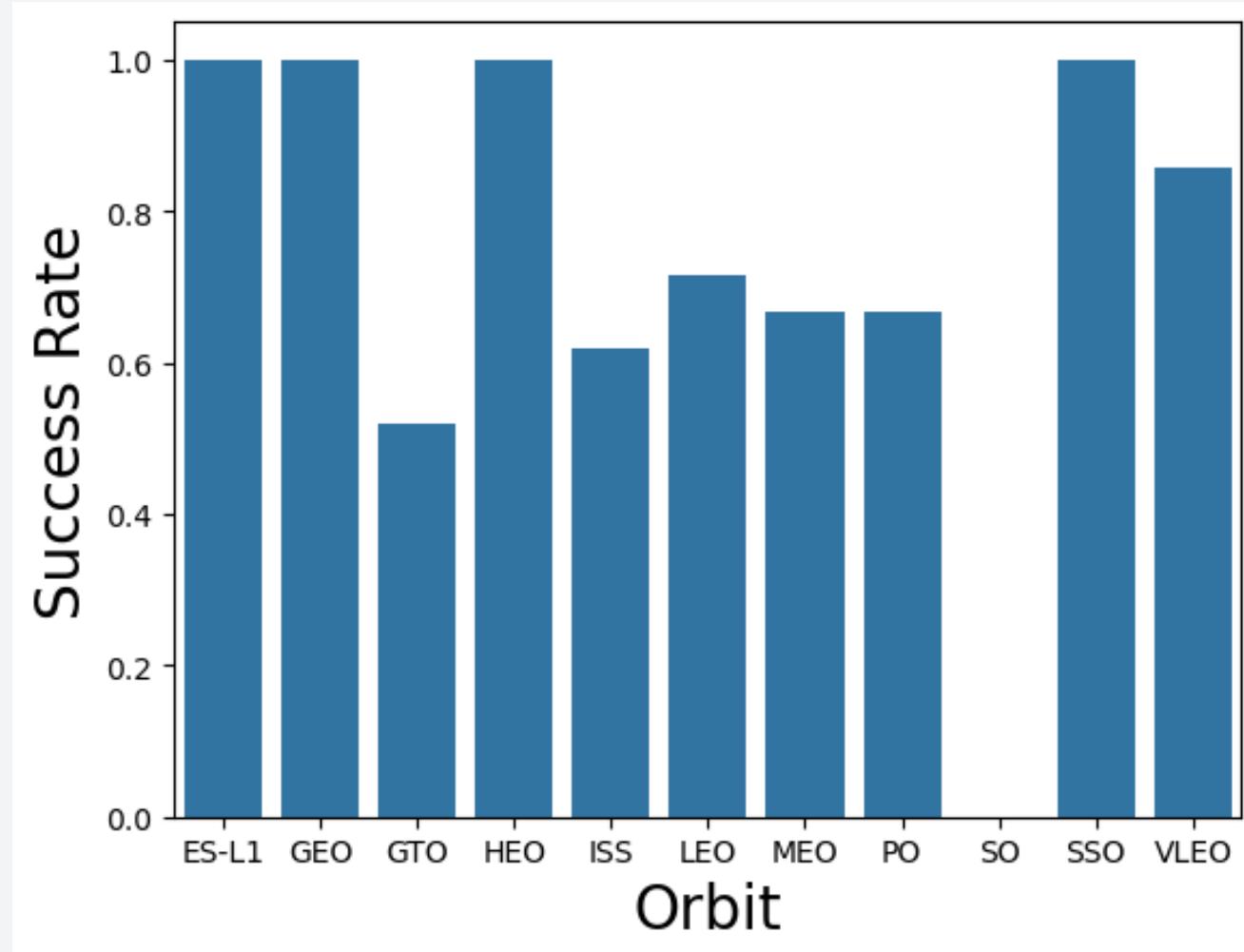
- ES-L1
- GEO
- HEO
- SSO

Успех в диапазоне 50-70%:

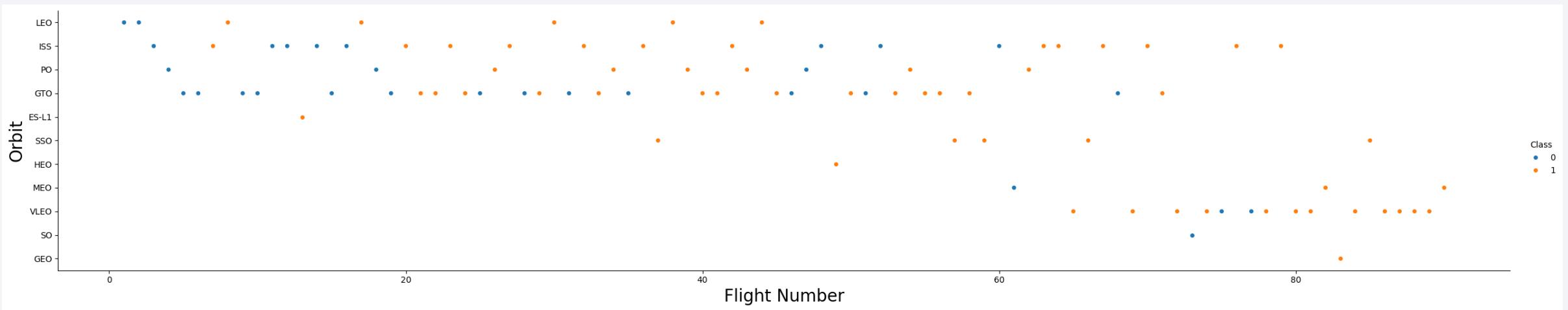
- VLEO
- LEO
- MEO
- PO
- ISS
- GTO

Уровень успеха 0%:

- SO



# Flight Number vs. Orbit Type



В случае низкой околоземной орбиты (LEO) уровень успешности кажется коррелирует с числом полетов, в то время как не наблюдается заметной связи между номером полета и уровнем успешности в геостационарной трансферной орбите (GTO).

# Payload vs. Orbit Type



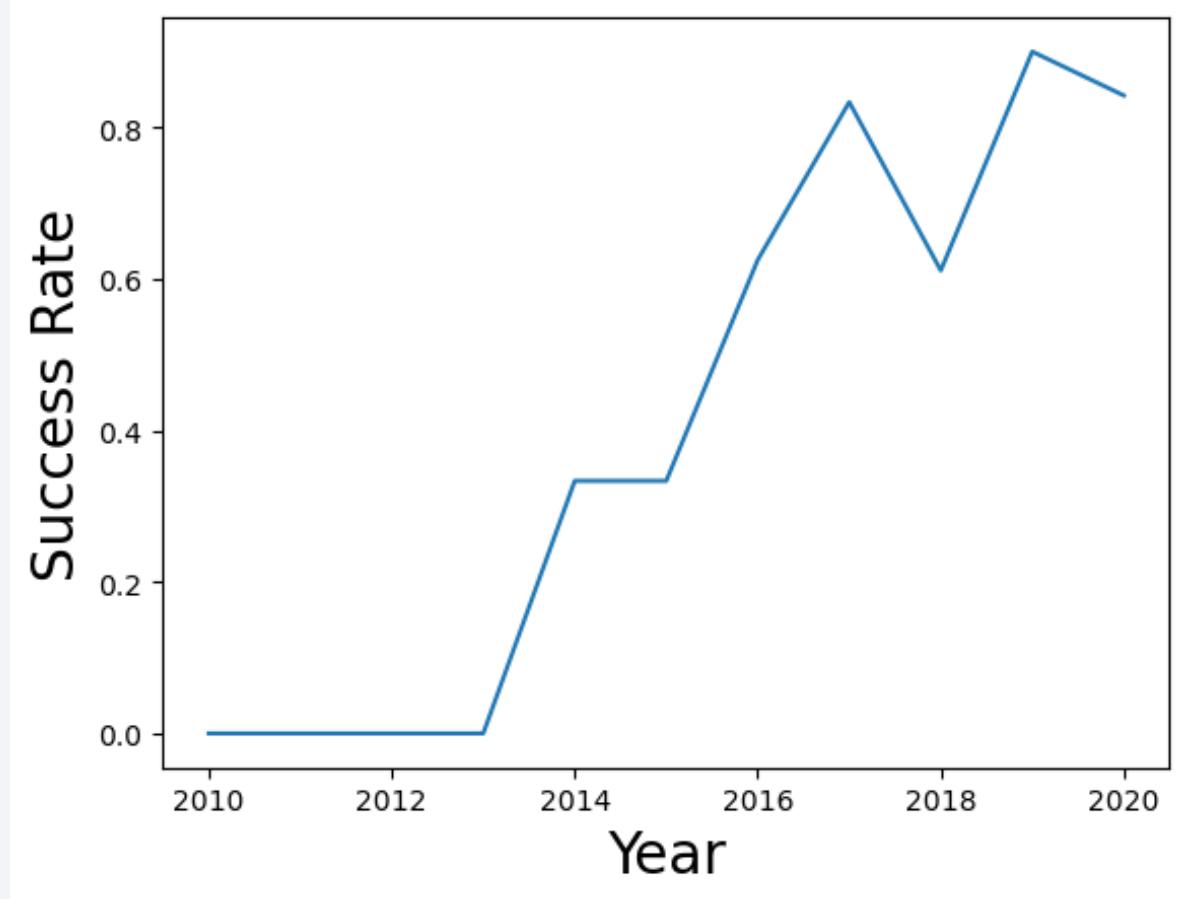
Для крупных грузов вероятность успешной или положительной посадки выше для миссий в полярной орбите, низкой околоземной орбите (LEO) и Международной космической станции (ISS). Однако различие между положительными и отрицательными посадками (неудачными миссиями) менее ясно для геостационарной трансферной орбиты (GTO), поскольку оба результата наблюдаются.

## Launch Success Yearly Trend

---

Вы можете заметить, что успешность запусков стабильно повышается с 2013 по 2017 год.

В 2018 году произошёл спад, а затем снова начался рост.



# Названия всех стартовых площадок

---

Стартовые площадки:

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

```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE;  
* sqlite:///my_data1.db  
Done.  


| Launch_Site  |
|--------------|
| CCAFS LC-40  |
| VAFB SLC-4E  |
| KSC LC-39A   |
| CCAFS SLC-40 |


```

Результат запроса предоставляет список площадок запуска, где проводились космические миссии. Среди них CCAFS LC-40 (Cape Canaveral Air Force Station), VAFB SLC-4E (Vandenberg Air Force Base) и KSC LC-39A (Kennedy Space Center).

# Названия стартовых площадок, начинающиеся с 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (F)
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 (F)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	N
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	N
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	N

Результат запроса перечисляет пять записей, где название площадки запуска начинается с "CCA". Все эти записи проходят с CCAFS LC-40 (Cape Canaveral Air Force Station, Launch Complex 40).

Они представляют различные космические миссии SpaceX, проведенные с этой площадки запуска.

# Общая полезная нагрузка

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS Total_Payload_Mass_NASA CRS FROM SPACEXTABLE WHERE Customer LIKE 'NASA (CRS)';
```

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

```
Done.
```

Total_Payload_Mass_NASA CRS
45596

Этот запрос вычисляет общую массу полезной нагрузки, перевезенной ракетами, запущенными NASA в рамках программы коммерческих поставок (CRS).

Общая масса полезной нагрузки составляет 45596 килограммов.

# Средняя масса полезной нагрузки F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS Average_Payload_Mass_F9_v1_1 FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1';  
* sqlite://my_data1.db  
Done.  
Average_Payload_Mass_F9_v1_1  
2928.4
```

- Этот запрос вычисляет среднюю массу полезной нагрузки, перевозимой ракетами версии F9 v1.1.
- Средняя масса полезной нагрузки составляет 2928,4 килограмма.

# Дата первой успешной посадки

```
%sql SELECT MIN(Date) AS First_Successful_Landing_Ground_Pad FROM SPACEXTABLE WHERE Landing_Outcome LIKE 'Success (ground pa
```

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

```
Done.
```

```
First_Successful_Landing_Ground_Pad
```

```
2015-12-22
```

Этот запрос определяет дату первого успешного исхода при посадке на землю.

Результат показывает, что первая успешная посадка на землю произошла 22 декабря 2015 года.

Это является важной вехой в истории технологии многоразовых ракет SpaceX.

Ракетоносители, которые успешно приземлились на автономном судне и имели массу полезной нагрузки между 4000 и 6000 килограмм

---

Имена ракетоносителей, которые успешно приземлились на автономном судне и имели массу полезной нагрузки более 4000, но менее 6000 килограмм:

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

```
%sql SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome LIKE 'Success (drone ship)' AND PAYLOAD_MASS_KG BETWEEN 4000 AND 6000;
* sqlite:///my_data1.db
Done.

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

Этот запрос перечисляет имена ракет-носителей, которые успешно приземлились на автономном судне и имели массу полезной нагрузки более 4000 килограмм, но менее 6000 килограмм.

# Общее число успешных и неудачных миссий

```
%sql SELECT Mission_Outcome, COUNT(*) AS Outcome_Count FROM SPACEXTABLE GROUP BY Mission_Outcome;
```

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

```
Done.
```

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

Этот запрос вычисляет общее количество успешных и неудачных результатов миссий на основе предоставленных данных.

Результат показывает, что было 98 успешных миссий и 1 неудача во время полета, а также 1 миссия с успехом, где статус полезной нагрузки неясен.

# Бустеры, перевозившие максимальную массу полезной нагрузки

Бустеры, которые перевозили максимальную массу полезной нагрузки:

```
[25]: %sql SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTA
* sqlite:///my_data1.db
Done.
[25]: 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
```

Эти бустеры перевозили наибольшие массы полезной нагрузки по сравнению с другими в предоставленном списке.

# Записи запусков 2015 года

---

В 2015 году произошли два неудачных приземления на барже:

- Месяц: Январь
- Исход приземления: Неудача (баржа-друган)
- Версия ракеты: F9 v1.1 B1012
- Место запуска: CCAFS LC-40

```
%sql SELECT SUBSTR(Date, 6, 2) AS Month, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE SUBSTR(Date, 0  
* sqlite:///my_data1.db  
Done.  


| 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 |


```

- Месяц: Апрель
- Исход приземления: Неудача (баржа-друган)
- Версия ракеты: F9 v1.1 B1015
- Место запуска: CCAFS LC-40

Эти случаи произошли во время запусков с комплекса 40 Военно-воздушной станции Кейп-Канаверал (CCAFS LC-40) в 2015 году.

# Ранжирование исходов приземления с 2010-06-04 по 2017-03-20

---

Ранжируя количество исходов приземления между 4 июня 2010 года и 20 марта 2017 года в порядке убывания, мы имеем:

- Отсутствует попытка: 10 случаев
- Успех (для автономного корабля): 5 случаев
- Неудача (для автономного корабля): 5 случаев
- Успех (для наземной площадки): 3 случая
- Контролируемый (океан): 3 случая
- Неконтролируемый (океан): 2 случая
- Неудача (парашют): 2 случая
- Предотвращено (для автономного корабля): 1 случай
- 

```
%sql SELECT Landing_Outcome, COUNT(*) AS Outcome_Count FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'  
* sqlite:///my_data1.db  
Done.
```

Landing_Outcome	Outcome_Count
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

The background of the slide is a photograph taken from space at night. It shows the curvature of the Earth against a dark blue and black void of space. City lights are visible as small white dots and larger clusters of light, primarily concentrated in the lower right quadrant where the United States appears. In the upper right, the green glow of the aurora borealis is visible in the atmosphere.

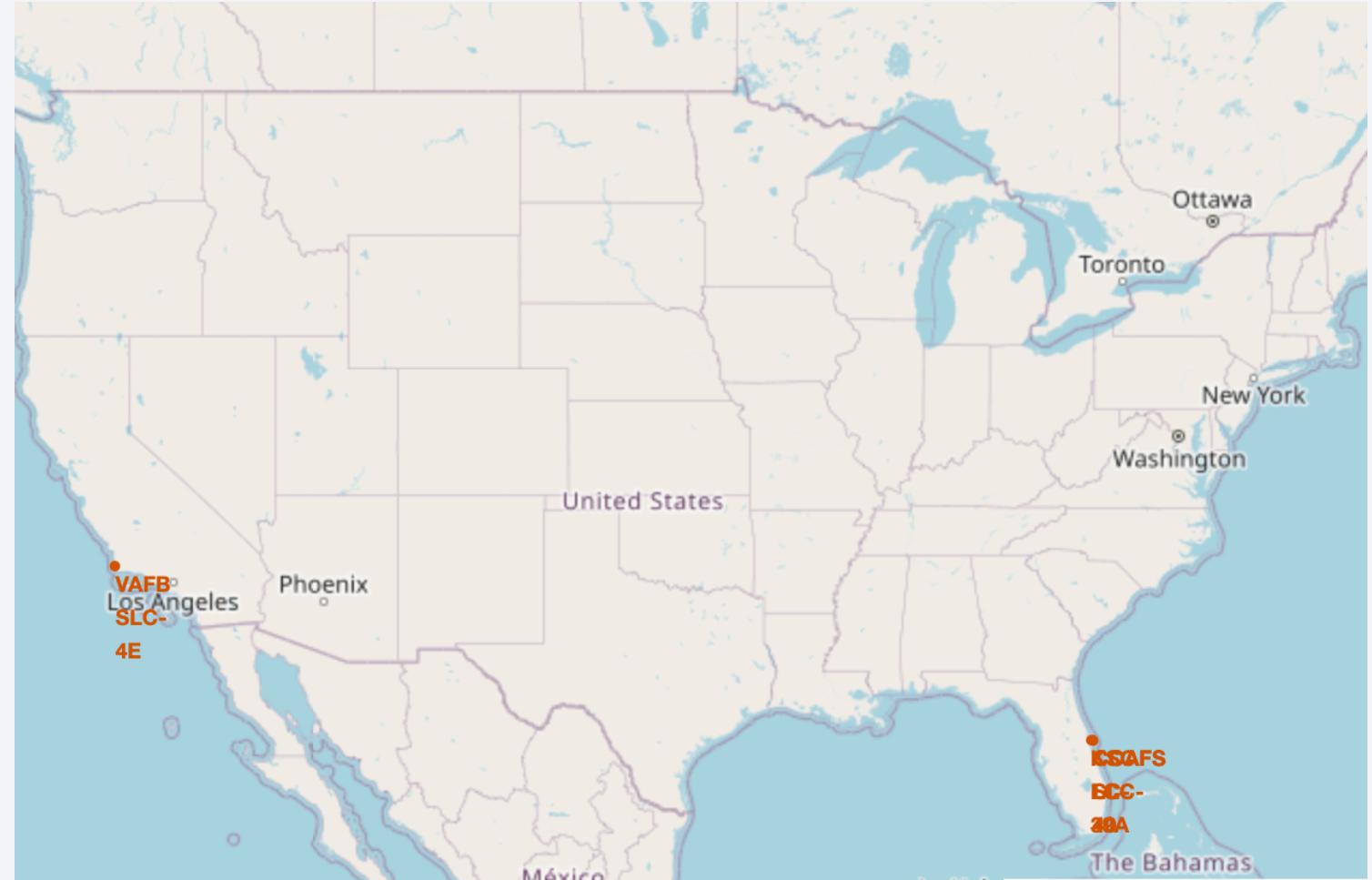
Section 3

# Launch Sites Proximities Analysis

# Карта со стартовыми площадками

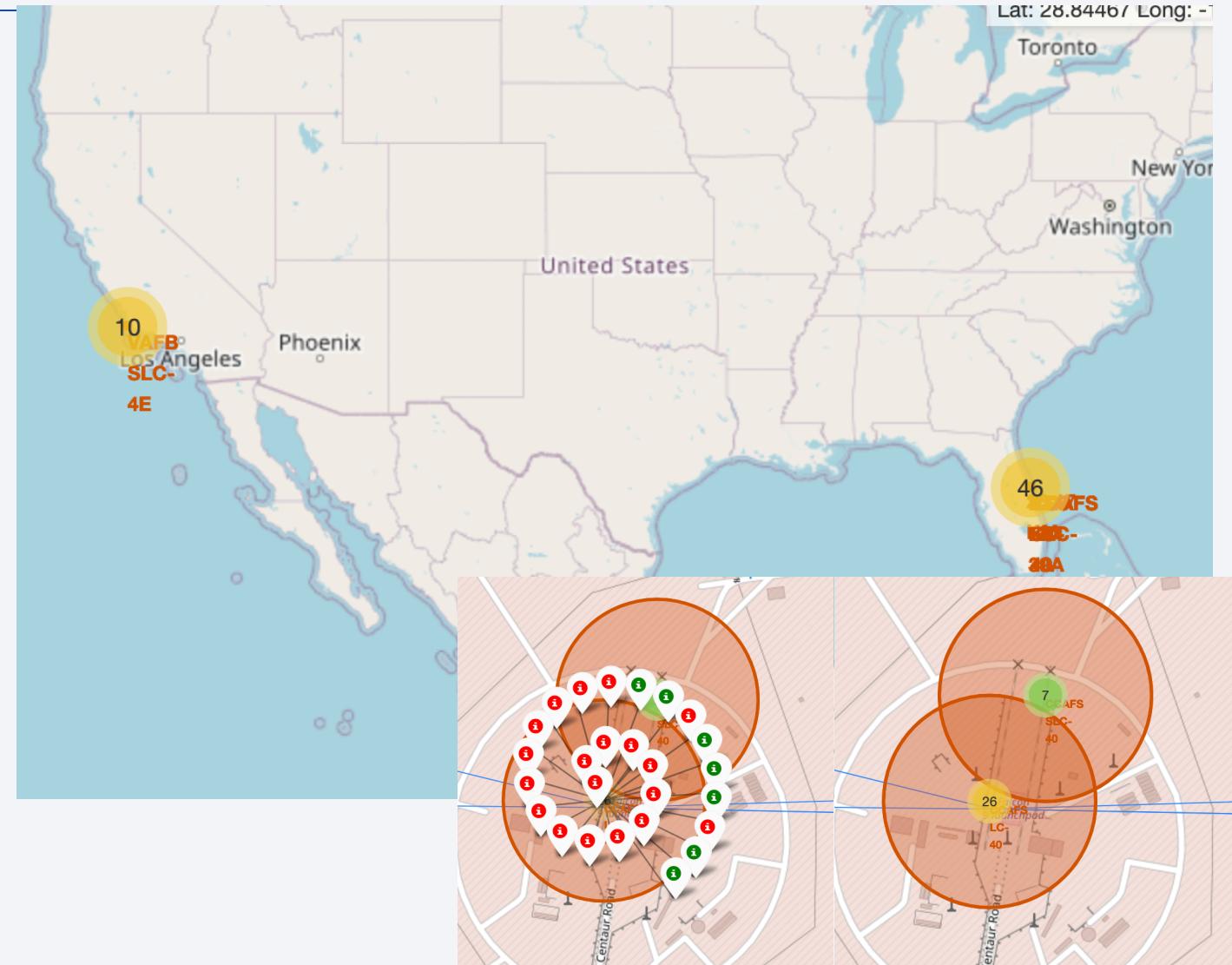
На снимке экрана показаны маркеры, обозначающие места запуска SpaceX в Соединенных Штатах. Эти площадки обычно расположены на военной базе BBC Кейп-Канаверал и космическом центре имени Кеннеди во Флориде.

- CCAFS LC-40,
- CCAFS SLC-40,
- KSC LC-39A,
- VAFB SLC-4E.



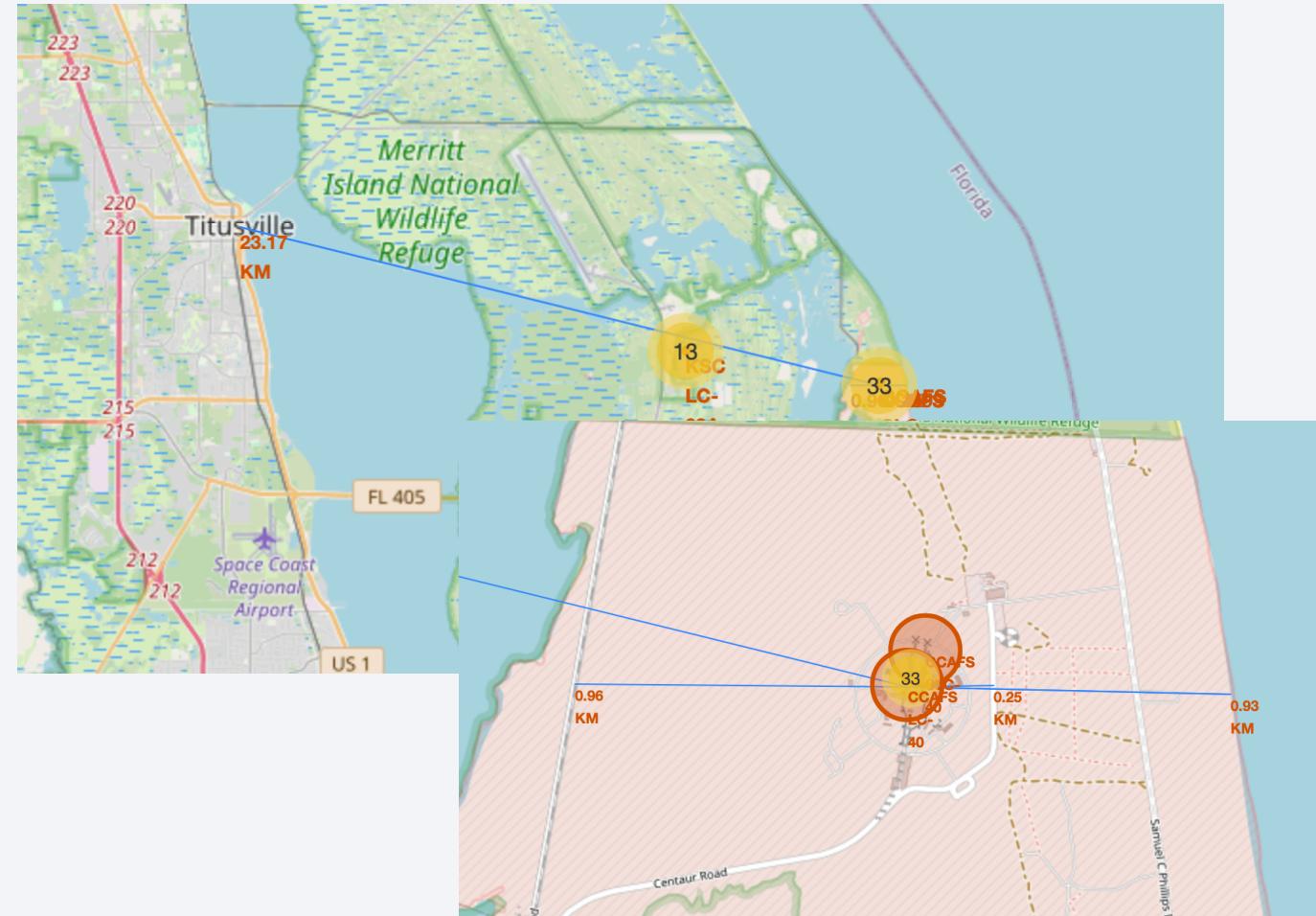
# Карта с цветными метками для результатов запуска

- Цветные метки результатов запусков: Снимок экрана иллюстрирует результаты запусков, отмеченные цветными метками на карте, где различные цвета обозначают различные исходы, такие как успех, неудача или другие категории.
- Метки места запуска: На карте отмечены места проведения  $10 + 46 = 56$  запусков. Эти метки обеспечивают пространственный контекст для распределения результатов запусков.
- Выделение фрагмента: Выделен определенный фрагмент карты, показывающий подмножество из 26 запусков, с 7 успешными и 19 неуспешными запусками.



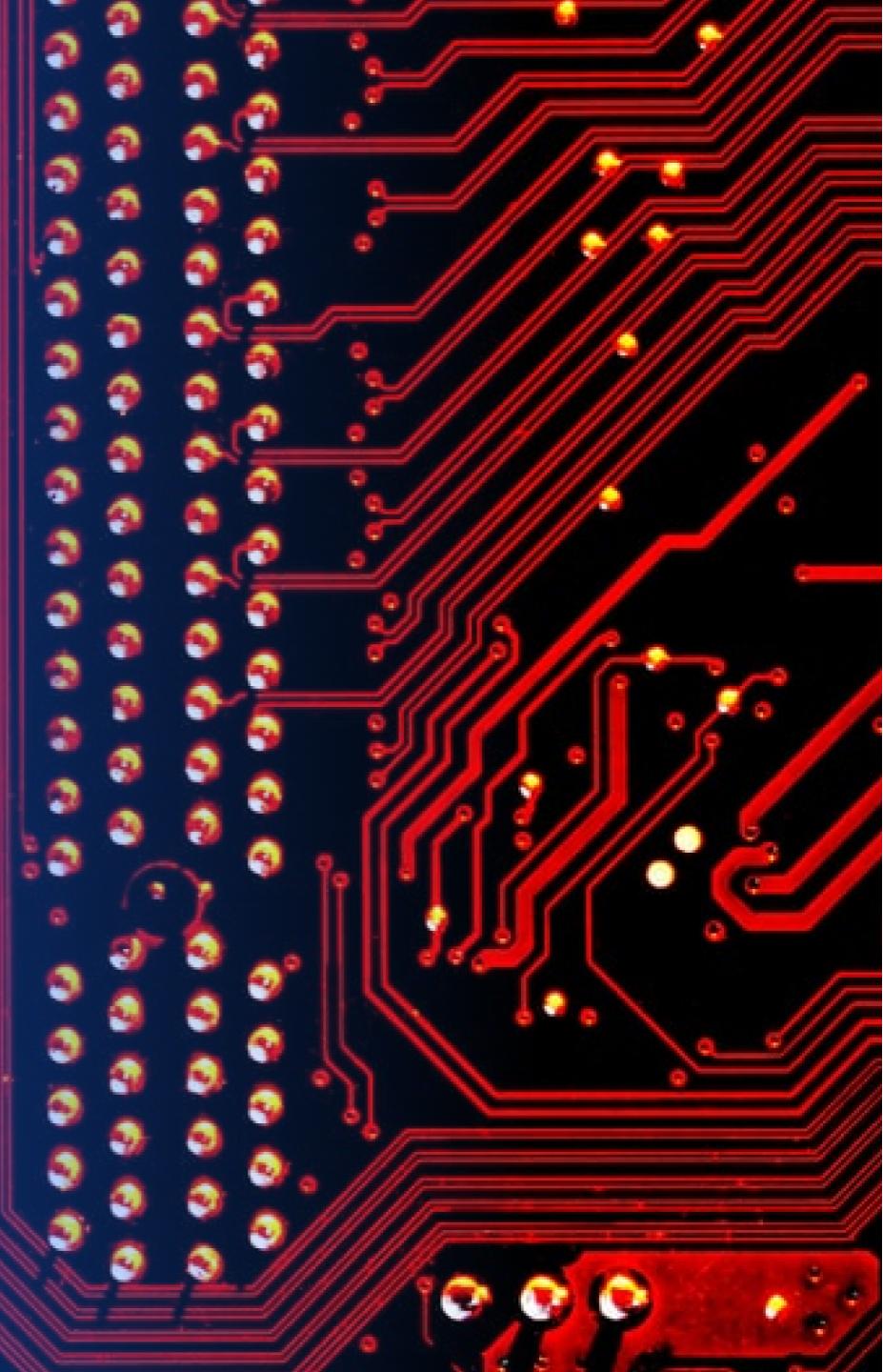
# Окрестности выбранного места запуска

- Маркер места запуска: Указывает точное местоположение выбранного места запуска.
- Маркеры близости: Выделяют близлежащие объекты, такие как железная дорога, шоссе и береговая линия.
- Вычисление расстояний: Отображает расстояния между местом запуска и близлежащими объектами, облегчая пространственное восприятие.
- Линия: Соединяет место запуска с каждым объектом близости, иллюстрируя прямые пути.
- Интерактивность: Позволяет пользователям изучать окружение места запуска, просматривать расстояния и получать пространственные представления.

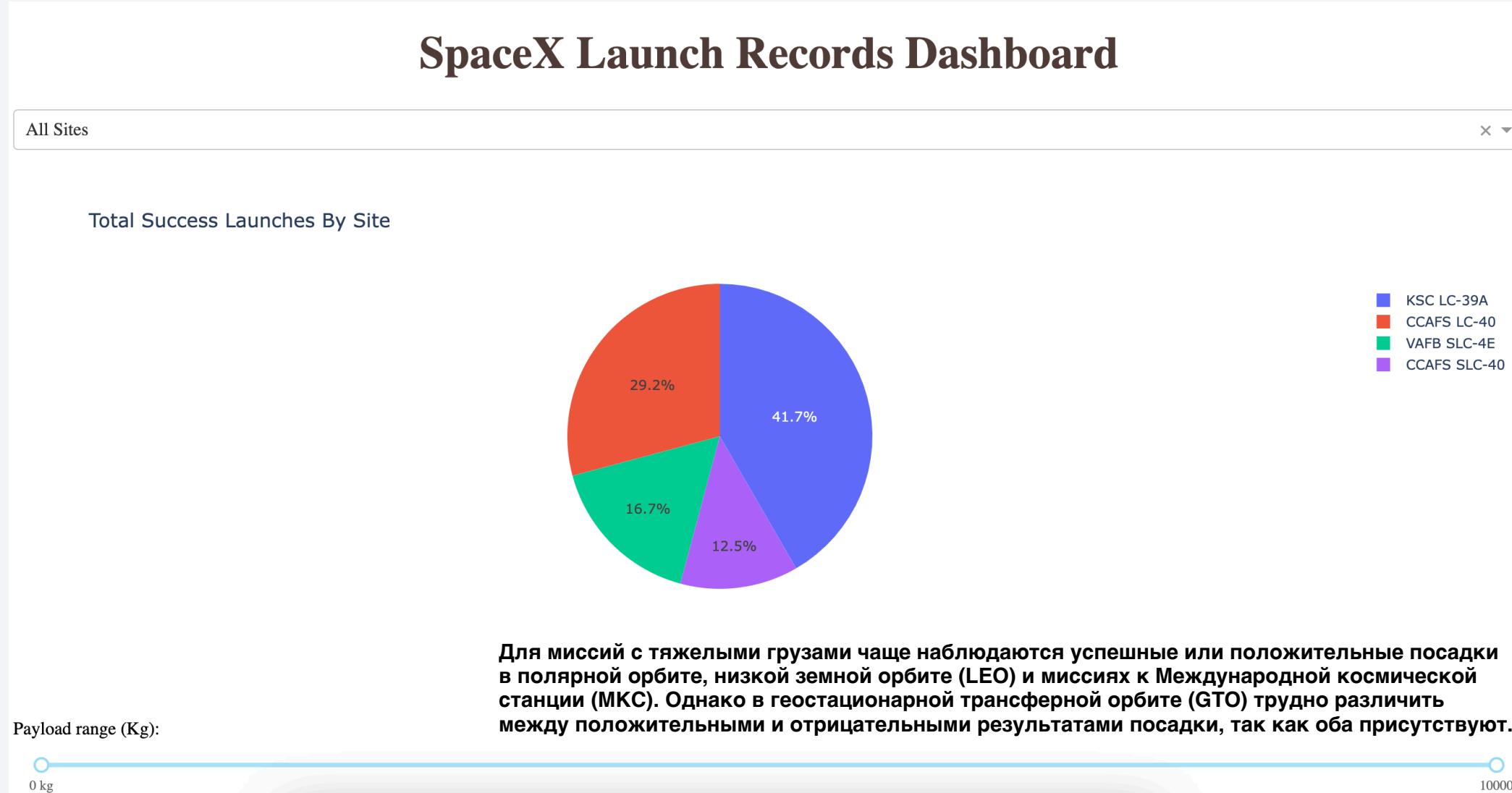


Section 4

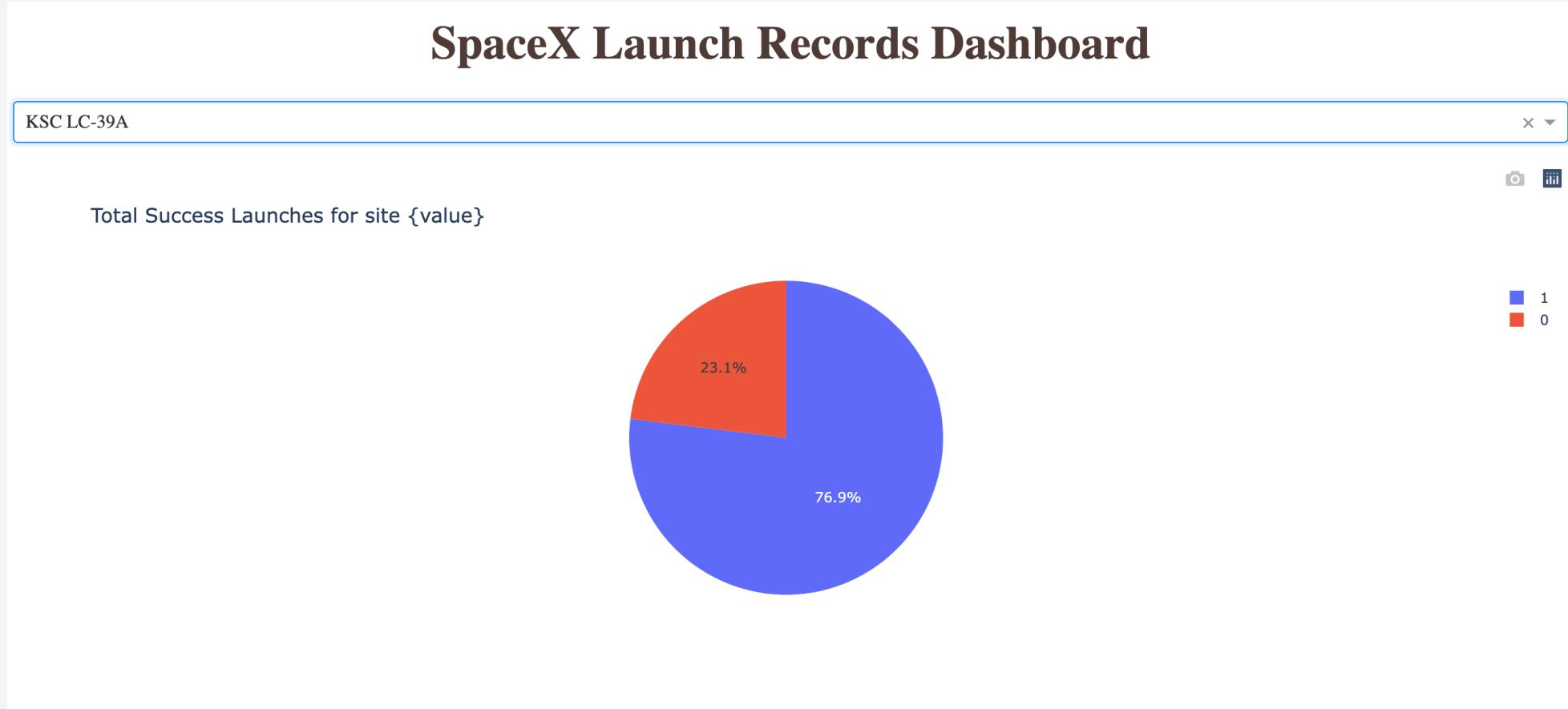
# Build a Dashboard with Plotly Dash



## Количество успешных запусков для всех площадок, в виде круговой диаграммы



## Круговая диаграмма для площадки с наивысшим коэффициентом успешных запусков



- Диаграмма в виде круга отображает коэффициенты успешных запусков для различных площадок, при этом KSC LC-39A имеет самый высокий коэффициент - 76,9%.

## Диаграмма рассеяния «Payload vs. Launch Outcome» для всех площадок

- Диаграмма рассеяния "Груз vs. Результат запуска" демонстрирует корреляцию между грузом и успешностью запуска на всех площадках, что позволяет анализировать, как различные диапазоны грузов и версии ракет-носителей влияют на успешность запусков.



Section 5

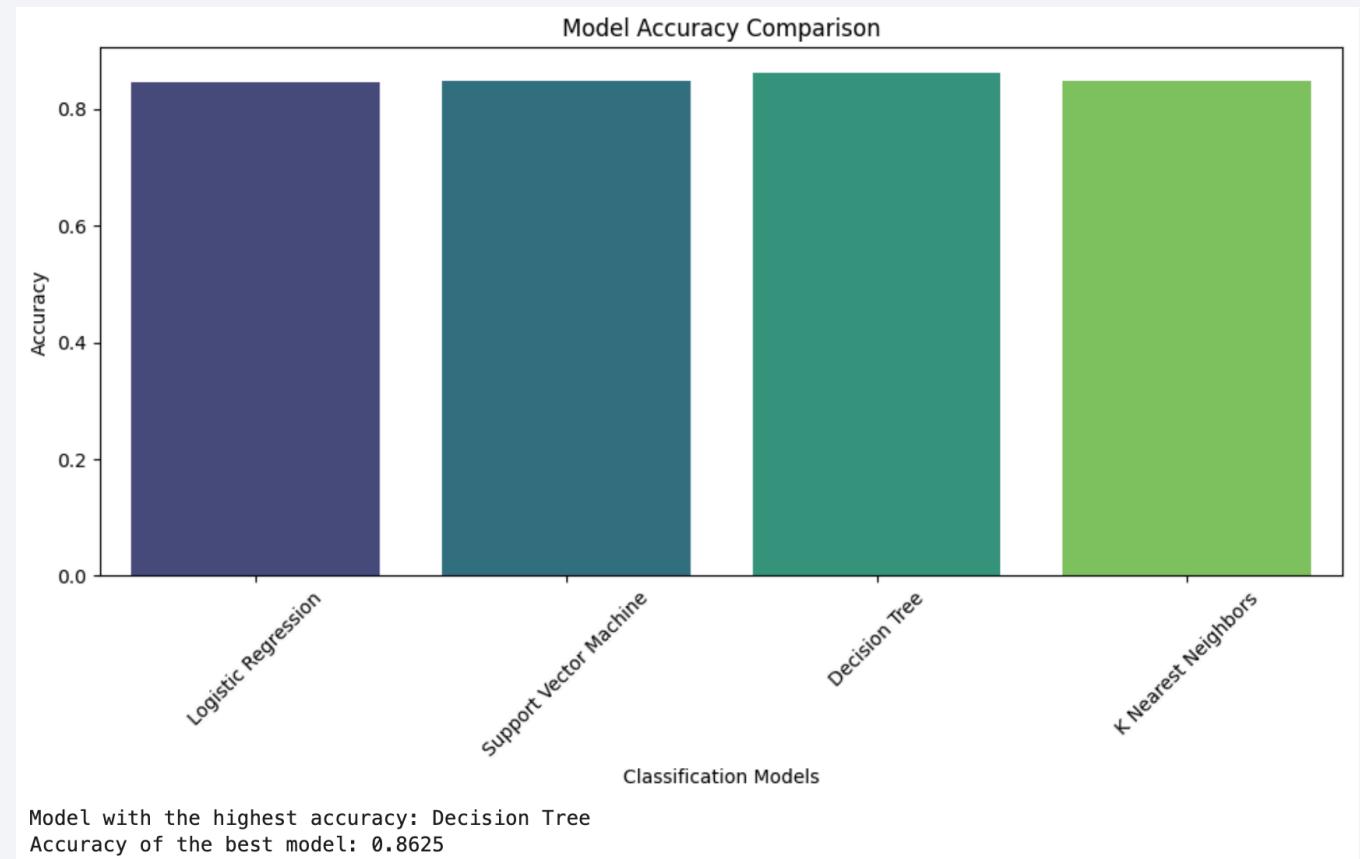
# Predictive Analysis (Classification)

# Точность классификации

- Результаты тестового набора не дают ясного понимания наиболее эффективной стратегии.
- Ограниченный размер тестовой выборки (18 образцов) может привести к схожим показателям в тестовом наборе.

Поэтому мы проанализировали все подходы, используя полный набор данных.

- После оценки всего набора данных модель решающего дерева выделилась как наиболее эффективный вариант, показав более высокие показатели и точность по сравнению с другими моделями.



# Матрица ошибок

- В этом случае модель правильно предсказала 11 из 12 положительных случаев ( $TP=11$ ), однако некоторые отрицательные случаи были ошибочно классифицированы как положительные ( $FP=4$ ). Кроме того, были 2 верно классифицированных отрицательных случая ( $TN=2$ ) и 1 ложноотрицательный случай ( $FN=1$ ).



## Выводы

---

Наш анализ данных SpaceX выявил ключевые факторы успешности миссий и ракет. Мы использовали сбор, очистку, визуализацию и моделирование данных.

Выводы показали, что количество запусков на космодроме коррелирует с его успешностью, и успешность запусков увеличивалась с 2013 по 2020 год.

Орбиты ES-L1, GEO, HEO, SSO и VLEO имели наибольший процент успешности, а космодром KSC LC-39A оказался самым успешным.

В качестве лучшего алгоритма машинного обучения для задачи был выбран классификатор дерева решений.

Thank you!

