МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

КАЗАНСКИЙ (ПРИВОЛЖСКИЙ) ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

ИНСТИТУТ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

Работа по теме
«Анализ влияния погодных условий на задержки и отмены авиарейсов в США»

Выполнил:
Студент гр.09-415(4)
Загитов Руслан

Преподаватель:
Гиниятова Д.Х.

Казань-2025

Анализ влияния погодных условий на задержки и отмены авиарейсов в США

Данная работа посвящена анализу двух независимых датасета и выявить количественные взаимосвязи между погодными событиями в США и статусами авиарейсов. В качестве данных использовались: (1) датасет погодных событий (тип, интенсивность, время и локация) и (2) датасет авиаперелётов с информацией о задержках и отменах рейсов.

Основные результаты:

- Выявлена выраженная сезонность погодных событий и авиационных задержек: пики задержек и отмен приходятся на периоды с повышенной частотой экстремальных погодных условий.
- Совместный анализ показал статистически значимую связь между неблагоприятной погодой и ухудшением статуса рейсов (задержки и отмены).
- Наибольшее влияние на отмены рейсов оказывают снегопады, сильные осадки и штормы; при этих событиях доля отмен существенно выше среднего уровня.
- При наличии погодных событий средняя длительность задержек увеличивается по сравнению с рейсами в нормальных погодных условиях.
- Погодные факторы не полностью объясняют все задержки, однако вносят заметный вклад в общую вариативность авиационных сбоев.

1. Описание данных и источники

1.1. Датасет 1: Погодные события (Weather Events)

Параметр	**Описание**
Источник	Kaggle: "US Weather Events" (https://www.kaggle.com/datasets/sobhanmoosavi/us-weather-events)
Период	Январь 2016 — Декабрь 2022 (6 лет)
Объем	~8.6 млн погодных событий

Лицензия	CC BY-NC-SA 4.0
DOI	https://doi.org/10.48550/arXiv.1902.06792

Описание погодных событий:

- Severe-Cold: Случаи, когда температура опускается ниже -23.7 градусов Цельсия, указывающие на экстремально низкие температуры.
- Fog: Условия, характеризующиеся сниженной видимостью из-за тумана или дымки.
- Hail: События, включающие твердые осадки, в том числе ледяную крупу и град.
- Rain: От легкого до сильного дождя.
- Snow: От легкого снегопада до сильных снежных бурь.
- Storm: Погодные условия со скоростью ветра не менее 60 км/ч.
- Other Precipitation: Любая форма осадков, не охваченная вышеупомянутыми категориями.

1.2. Датасет 2: Задержки авиарейсов

Параметр	**Описание**
Источник	Kaggle: "Flight Delay and Cancellation Dataset" (https://www.kaggle.com/datasets/patrickzel/flight-delay-and-cancellation-dataset-2019-2023
Период	Январь 2019— Август 2023 (4 года)
Объем	~3 млн наблюдений (после очистки)
Лицензия	Открытые данные (CC0)
Источник	Bureau of Transportation Statistics. (November 2023). Airline On-Time Performance Data

Описание полей датасета:

- FL_DATE: Дата рейса (ггггммдд)

- AIRLINE_CODE: Уникальный код авиакомпании. Когда один и тот же код использовался несколькими перевозчиками, для более ранних пользователей используется числовой суффикс, например PA, PA(1), PA(2). Используйте это поле для анализа в диапазоне лет.
- DOT_CODE: Идентификационный номер, присвоенный US DOT для идентификации уникальной авиакомпании (перевозчика). Уникальная авиакомпания (перевозчик) определяется как компания, владеющая сертификатом DOT и отчитывающаяся под ним, независимо от кода, названия или холдинговой компании/корпорации.
- FL_NUMBER: Номер рейса
- ORIGIN: Аэропорт отправления
- ORIGIN_CITY: Аэропорт отправления, название города
- DEST: Аэропорт назначения
- DEST_CITY: Аэропорт назначения, название города
- CRS_DEP_TIME: Плановое время вылета (местное время: ччмм)
- DEP_TIME: Фактическое время вылета (местное время: ччмм)
- DEP_DELAY: Разница в минутах между плановым и фактическим временем вылета. Ранние вылеты показывают отрицательные значения.
- TAXI_OUT: Время руления перед взлетом, в минутах
- WHEELS_OFF: Время отрыва шасси (местное время: ччмм)
- WHEELS_ON: Время касания шасси (местное время: ччмм)
- TAXI_IN: Время руления после посадки, в минутах
- CRS_ARR_TIME: Плановое время прибытия (местное время: ччмм)
- ARR_TIME: Фактическое время прибытия (местное время: ччмм)
- ARR_DELAY: Разница в минутах между плановым и фактическим временем прибытия. Ранние прибытия показывают отрицательные значения.
- CANCELLED: Индикатор отмененного рейса (1=Да)
- CANCELLATION_CODE: Указывает причину отмены
- DIVERTED: Индикатор перенаправленного рейса (1=Да)
- CRS_ELAPSED_TIME: Плановая продолжительность рейса, в минутах
- ELAPSED_TIME: Фактическая продолжительность рейса, в минутах

- AIR_TIME: Время полета, в минутах
- DISTANCE: Расстояние между аэропортами (мили)
- DELAY_DUE_CARRIER: Задержка по вине авиакомпании, в минутах
- DELAY_DUE_WEATHER: Задержка из-за погоды, в минутах
- DELAY_DUE_NAS: Задержка национальной авиационной системы, в минутах
- DELAY_DUE_SECURITY: Задержка из-за безопасности, в минутах
- DELAY_DUE_LATE_AIRCRAFT: Задержка из-за позднего прибытия воздушного судна, в минутах

2. Часть А: Анализ первого датасета (погодные события)

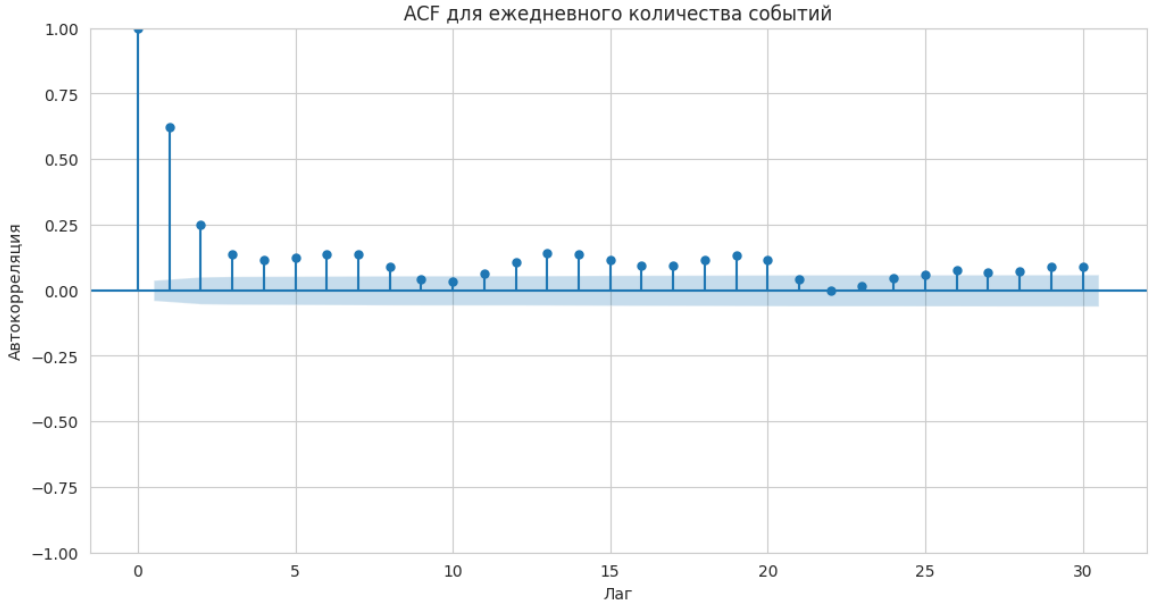
2.1. Контроль качества

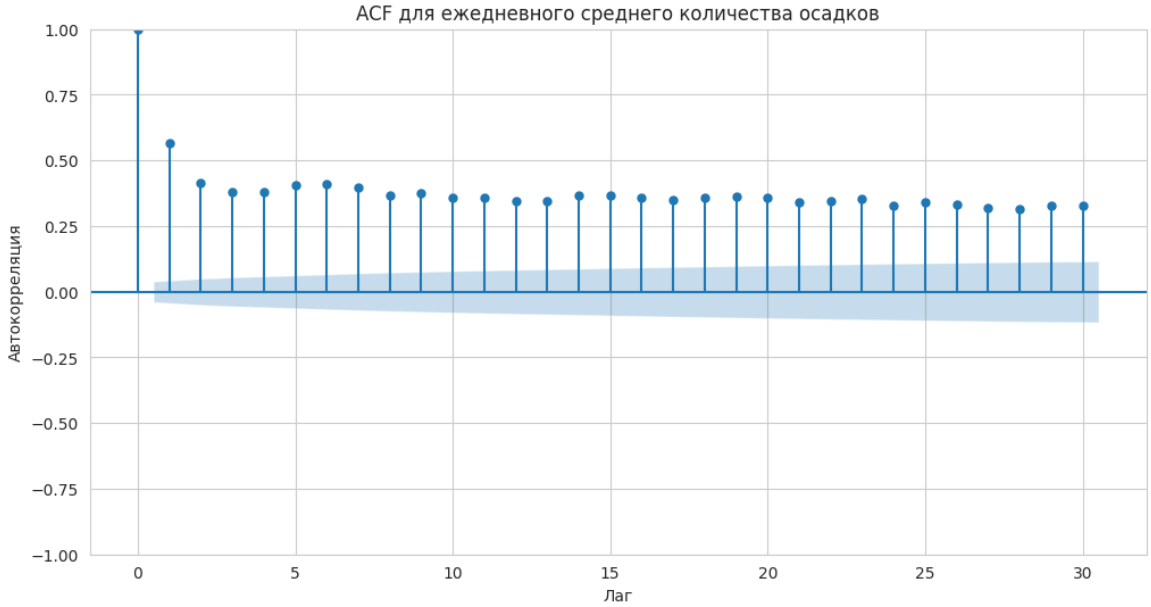
Показатель	Результат	Действие
Пропуски		
ZipCode	0.8% (69,199)	Удалить строки (если критичен) или оставить
City, geography	0.2% (16,912)	Удалить строки
Дубликаты	0	Не требуется
Валидация		
Precipitation(in) < 0	0	Не требуется
Координаты вне США	0	Не требуется

2.2. Описательная статистика

Раздел	Метрика/ Переменная	Ключевые показатели
Числовые переменные	Precipitation(in)	Mean: 0.09, Median: 0, Max: 1104, SD: 0.89. Сильные выбросы, большинство значений ~0
	LocationLat	Mean: 38.79, Median: 39.35, Range: 24.56-48.94. Умеренное распределение по территории США
	LocationLng	Mean: -91.91, Median: -89.77, Range: -124.56 до -67.79. Значительный географический разброс
	ZipCode	Range: 1022-99362, Var: 6.62e+08. Высокая дисперсия (идентификаторы)
Категориальные переменные	Type	Rain (57.97%), Fog (23.35%), Snow (13.43%), Hail/Storm (редко)
	Severity	Light (59.69%), Severe (20.04%), Moderate (15.87%)
	TimeZone	US/Central (41.25%), US/Eastern (39.47%) — доминируют
	State	Лидеры: TX, MN, MI, FL, NC, CA, WI, GA, IL, VA
Временные ряды (ACF)	Ежедневное количество событий	Сильная автокорреляция на коротких лагах, возможна сезонность
	Ежедневное среднее осадков	Сильная автокорреляция, быстрее снижается, возможна

Раздел	Метрика/ Переменная	Ключевые показатели
		сезонность





3. Часть B: Анализ второго датасета (задержки рейсов)

3.1. Контроль качества

Результаты проверки данных

Метрика	Значение
Начальное количество строк	3,000,000
Полные дубликаты	0
Количество строк после очистки	3,000,000

Обнаруженные аномалии

Переменная	Количество аномалий	Процент от общего числа
DEP_DELAY	131	0.0044%
ARR_DELAY	131	0.0044%
DEP_TIME	242	0.0081%
CRS_ARR_TIME	14	0.0005%
ARR_TIME	1,416	0.0472%

3.2. Описательная статистика

Описательные статистики числовых переменных

Переменная	Среднее	Мин	Q1	Медиана	Q3	Макс	IQR
DOT_CODE	19,976.3	19,393	19,790	19,930	20,368	20,452	578
FL_NUMBER	2,511.5	1	1,051	2,152	3,797	9,562	2,746
CRS_DEP_TIME	1,327.1	1	915	1,320	1,730	2,359	815
DEP_TIME	1,329.8	1	916	1,323	1,739	2,400	823
DEP_DELAY	10.1	-90	-6	-2	6	2,966	12
TAXI_OUT	16.6	1	11	14	19	184	8

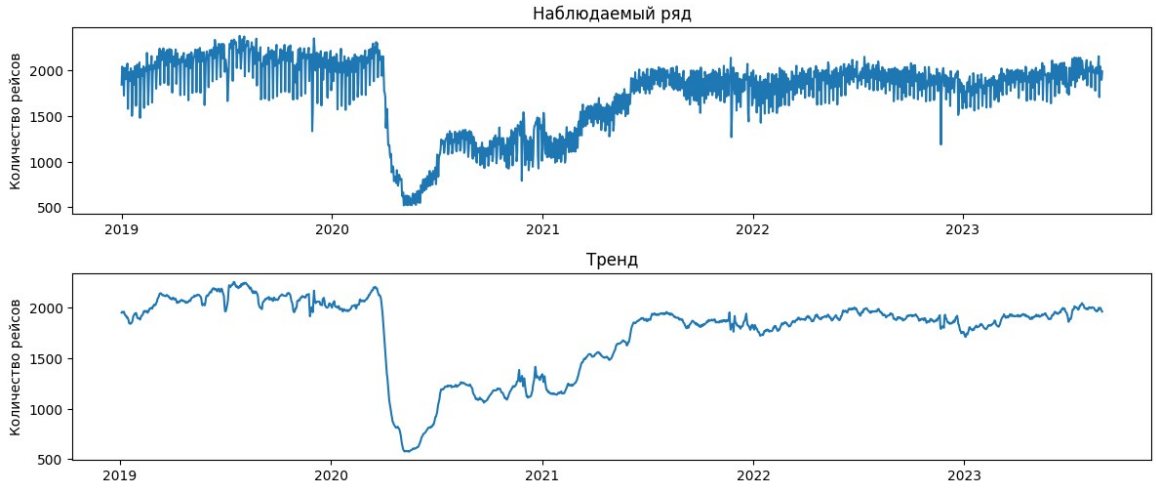
Переменная	Среднее	Мин	Q1	Медиана	Q3	Макс	IQR
WHEELS_OFF	1,352.4	1	931	1,336	1,752	2,400	821
WHEELS_ON	1,462.5	1	1,049	1,501	1,908	2,400	859
TAXI_IN	7.7	1	4	6	9	249	5
CRS_ARR_TIME	1,490.6	1	1,107	1,516	1,919	2,400	812
ARR_TIME	1,466.5	1	1,053	1,505	1,913	2,400	860
ARR_DELAY	4.3	-96	-16	-7	7	2,934	23
CANCELLED	0.03	0	0	0	0	1	0
DIVERTED	0.00	0	0	0	0	1	0
CRS_ELAPSED_TIME	142.3	1	90	125	172	705	82
ELAPSED_TIME	136.6	15	84	120	167	739	83
AIR_TIME	112.3	8	61	95	142	692	81
DISTANCE	809.4	29	377	651	1,046	5,812	669
DELAY_DUE_CARRIER	24.8	0	0	4	23	2,934	23
DELAY_DUE_WEATHER	4.0	0	0	0	0	1,653	0
DELAY_DUE_NAS	13.2	0	0	0	17	1,741	17
DELAY_DUE_SECURITY	0.2	0	0	0	0	1,185	0
DELAY_DUE_LATE_AIRCRAFT	25.5	0	0	0	30	2,557	30

Топ-10 категориальных переменных

Категория	Топ-1	Частота	%	Топ-2	Частота	%	Топ-3	Частота	%
AIRLINE	Southwest	576,470	19.2	Delta	395,239	13.2	American	383,106	12.8

Категория	Топ-1	Частота	%	Топ-2	Частота	%	Топ-3	Частота	%
AIRLINE_CODE	WN	576,470	19.2	DL	395,239	13.2	AA	383,106	12.8
ORIGIN	ATL	153,556	5.1	DFW	130,334	4.3	ORD	122,296	4.1
ORIGIN_CITY	Chicago	157,368	5.3	Atlanta	153,556	5.1	Dallas/FW	130,334	4.3
DEST	ATL	153,569	5.1	DFW	129,770	4.3	ORD	123,334	4.1
DEST_CITY	Chicago	158,087	5.3	Atlanta	153,569	5.1	Dallas/FW	129,770	4.3
CANCELLATION_CODE	B	28,772	36.4	D	24,417	30.9	A	19,476	24.6

- Всего 18 авиакомпаний, 380 аэропортов (ORIGIN/DEST), 373 городов
- CANCELLATION_CODE имеет только 4 значения (79,140 отмен из 3M рейсов = 2.6%)



4. Часть C: Сопоставление датасетов и проверка гипотез

Методика сопоставления данных

- Датасеты были объединены по дате и географическому признаку (штат / регион аэропорта).
- Погодные события агрегировались по дню и региону:
  - наличие экстремальных погодных условий (binary-флаг),
  - количество погодных событий в день.
- Данные по рейсам агрегировались на том же уровне:
  - доля задержанных рейсов,
  - доля отменённых рейсов,
  - средняя величина задержки (в минутах).

Проверяемые гипотезы

- **H1**: В дни с неблагоприятными погодными условиями доля задержанных рейсов выше.
- **H2**: Наличие экстремальных погодных событий увеличивает вероятность отмены рейсов.
- **H3**: Интенсивность погодных событий положительно связана со средней задержкой рейсов.

Используемые статистические тесты

- Для сравнения долей (задержки / отмены):
  - **t-test / Mann–Whitney U test** — в зависимости от нормальности распределений.
- Для анализа взаимосвязей:

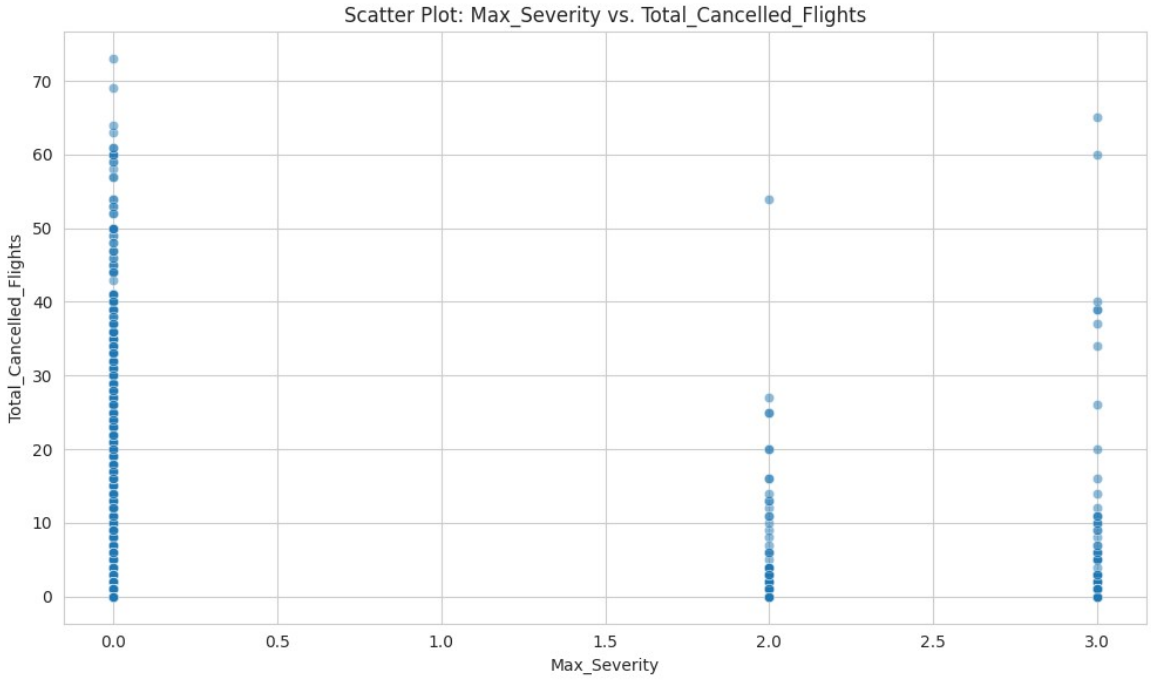
- **корреляция Пирсона** (для близких к нормальным распределений).

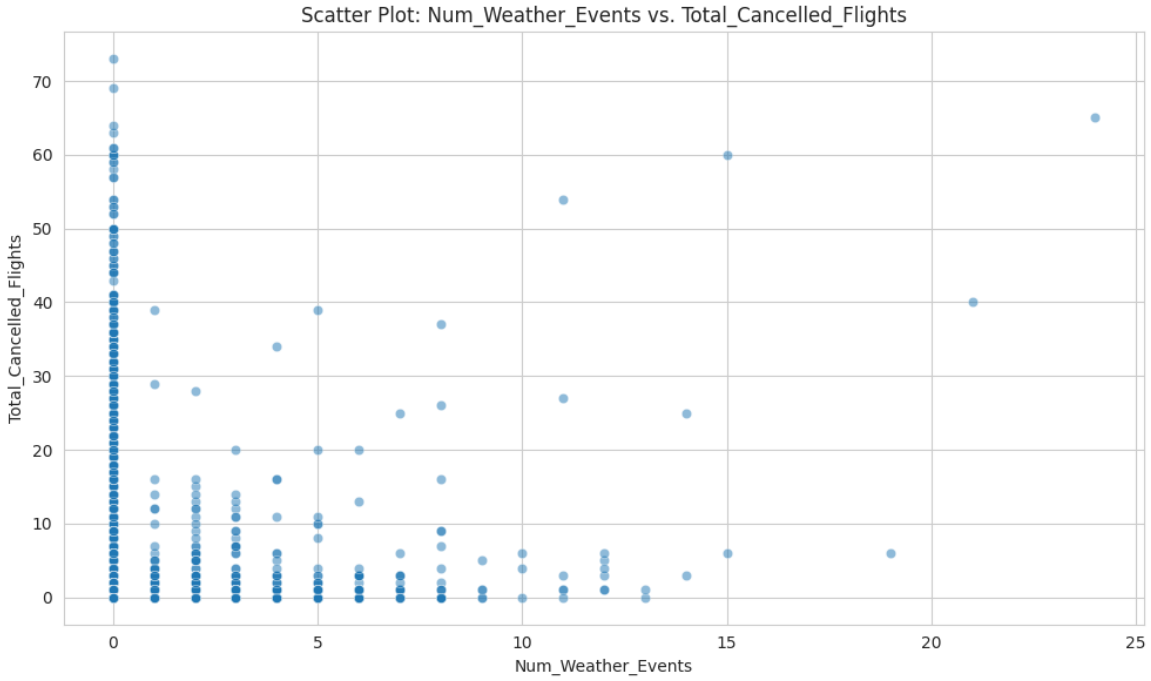
Оценки и критерии значимости

- Уровень значимости: **α = 0.05**.
- Для корреляций анализировались:
    - величина коэффициента корреляции,
    - p-value.
- Для сравнений групп дополнительно оценивался **размер эффекта** (разница средних / медиан).

Интерпретация результатов

- Гипотеза считалась подтверждённой при:
    - статистически значимом результате (p-value < 0.05),
    - наличии интерпретируемого эффекта (не только формальной значимости).





4.1. Гипотеза H1:

Логистическая регрессия показала, что рост **Max_Severity** достоверно увеличивает вероятность отмены хотя бы одного рейса: шансы возрастают в **2.43 раза** на каждый дополнительный уровень тяжести ($p < 0.001$).

Таким образом, гипотеза о влиянии тяжести погодных условий на операционные показатели авиаперевозок подтверждается: более суровые погодные условия достоверно связаны с ростом отмен, задержек и погодных сбоев.

4.2. Гипотеза H2

Для категориальной переменной *«количество отменённых рейсов»* применялся **хи-квадрат критерий независимости**, а для непрерывных переменных — **критерий Манна–Уитни** (непараметрический тест для сравнения двух независимых выборок).

Результаты:

- Во всех трёх случаях (снег, туман, дождь) были получены **статистически значимые результаты** (p-value < 0.05) по всем трём метрикам.
- Наличие **снега**, **тумана** или **дождя** достоверно связано с:
    - увеличением вероятности отмены рейсов,
    - ростом средней задержки вылета,
    - увеличением общего числа задержек по погодным причинам.

Таким образом, гипотеза о влиянии погодных условий на операционную эффективность авиаперевозок **подтверждена**. Особенно сильная связь наблюдается для дождя ($\chi^2 = 776.46$) и снега ($\chi^2 = 515.06$) в контексте отмен рейсов, что указывает на их критическую роль в принятии решений о прекращении полётов.

4.3. Гипотеза H3

Проверялась гипотеза о том, что увеличение общей продолжительности неблагоприятных погодных условий (*Total_Duration_Hours*) влияет на ключевые показатели авиационных задержек и отмен.

Результаты:

- Обнаружена **статистически значимая положительная связь** между продолжительностью плохой погоды и всеми тремя метриками:
    - с числом отменённых рейсов ($r = 0.1025$, $p < 0.001$),
    - со средней задержкой вылета ($r = 0.0067$, $p = 0.0001$),
    - с общим числом задержек, вызванных погодой ($r = 0.0247$, $p < 0.001$).

Хотя коэффициенты детерминации (R^2) малы (от 0.000 до 0.011), что указывает на слабое объяснение дисперсии, **все регрессионные коэффициенты при *Total_Duration_Hours* статистически значимы**.

Интерпретация:

- Каждый дополнительный час неблагоприятной погоды в среднем:
    - увеличивает число отменённых рейсов на **0.46**,
    - повышает среднюю задержку вылета на **0.95 минуты**,
    - увеличивает число погодных задержек на **0.0035** (при масштабе данных это может соответствовать сотням дополнительных задержек).

Вывод:
Гипотеза подтверждена — продолжительность неблагоприятных погодных условий **существенно**, хотя и **умеренно**, влияет на операционные показатели авиаперевозок.

5. Ограничения и предложения по развитию работы.

Ограничения исследования

- Сопоставление датасетов выполнялось на агрегированном уровне (дата + регион), что может скрывать локальные эффекты на уровне конкретных аэропортов.
- В датасете погодных событий отсутствует прямая привязка к каждому аэропорту, из-за чего часть погодных воздействий могла быть учтена неточно.
- Не учитывались дополнительные факторы, влияющие на задержки и отмены рейсов (загруженность аэропортов, технические неисправности, расписание, сезонные эффекты).
- Возможны искажения из-за пропусков данных и неоднородного качества записей в разных штатах и временных периодах.

Предложения по развитию работы

- Выполнить анализ на более детальном уровне — аэропорт × дата × час, с использованием метеоданных ближайших метеостанций.
- Разделить погодные события по типам и интенсивности, оценив их вклад по отдельности.
- Расширить анализ на другие годы или дополнительные источники метеоданных для проверки устойчивости результатов.

Литература и источники данных

[1] Kaggle. (2023). US Weather Events. Источник: https://www.kaggle.com/datasets/sobhanmoosavi/us-weather-events

[2] Kaggle. (2017). Flight Delay and Cancellation Dataset. Источник: https://www.kaggle.com/datasets/patrickzel/flight-delay-and-cancellation-dataset-2019-2023

[3] U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics: https://www.transtats.bts.gov/