ОТЧЕТ О ПРОЕКТЕ: АНАЛИТИЧЕСКАЯ ПЛАТФОРМА ДЛЯ СНИЖЕНИЯ ДОРОЖНО-ТРАНСПОРТНЫХ ПРОИСШЕСТВИЙ

Цель проекта: создание аналитической платформы для анализа дорожно-транспортных происшествий в США с целью выявления факторов риска, разработки превентивных мер и снижения аварийности.

Бизнес-проблема: ежегодно в США происходят миллионы дорожно-транспортных происшествий, которые приводят к человеческим жертвам, материальному ущербу и экономическим потерям. Отсутствие системного анализа факторов, влияющих на аварийность, ограничивает возможности для разработки эффективных профилактических мер.

Ключевые задачи:

1. Сбор и обработка данных о 7.7+ миллионах ДТП;
2. Выявление временных, географических и погодных паттернов;
3. Оценка качества данных и их пригодности для анализа;
4. Формулирование конкретных рекомендаций для повышения безопасности.

Методология: полный ETL-цикл: извлечение, преобразование, загрузка, анализ.

1. Исходные данные

Таблица 1 – Исходные данные.

Параметр	Значение	Примечание
Объем данных	7,728,394 записей	Данные по всем штатам США
Количество признаков	46 колонок	География, время, погода, тяжесть
Размер файла	3.06 GB	CSV формат
Период	2016-2023 гг.	8 лет наблюдений
Источник	US Accidents Dataset	Публичный датасет

2. ETL процесс и обработка данных

2.1. Extract (Извлечение)

- Автоматическая загрузка с Google Drive;
- Поддержка различных форматов ссылок;
- Прогресс-бар загрузки.

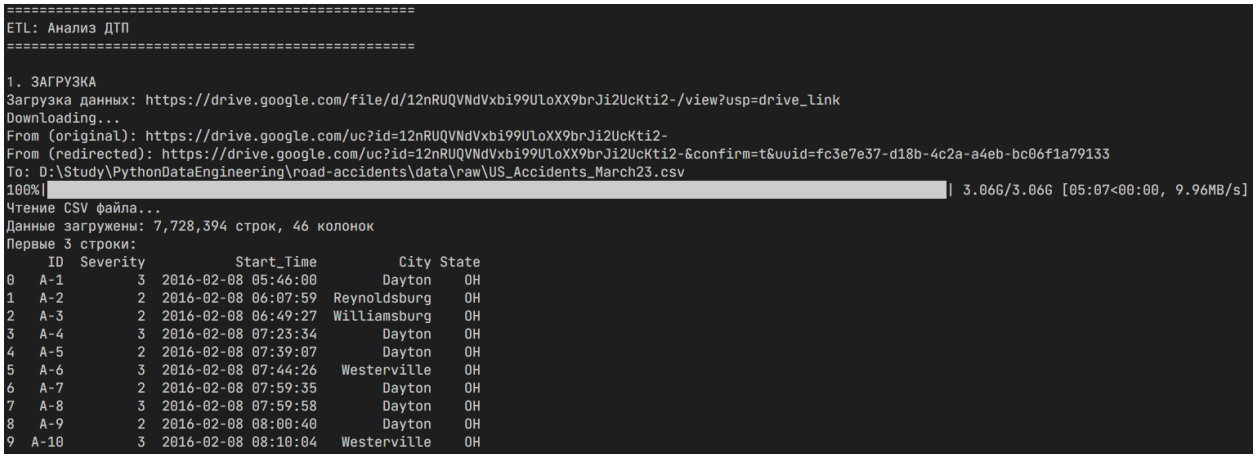

Рисунок 1 – Процесс загрузки данных с Google Drive.

2.2. Transform (Преобразование)

Таблица 2 – Обработка временных данных.

Действие	Результат
Конвертация Start_Time/End_Time	datetime формат
Создание производных признаков	Year, Month, Day, Hour, Weekday
Общее количество новых признаков	5

Таблица 3 – Типизация данных

Тип данных	Количество колонок	Примеры
datetime64[ns]	2	Start_Time, End_Time
int64	1	Severity
float64	17	Температура, влажность, скорость ветра
bool	13	Флаги условий (перекрестки, светофоры)
category	7	Штат, город, часовой пояс, погода
object	14	ID, описание, улица, почтовый индекс

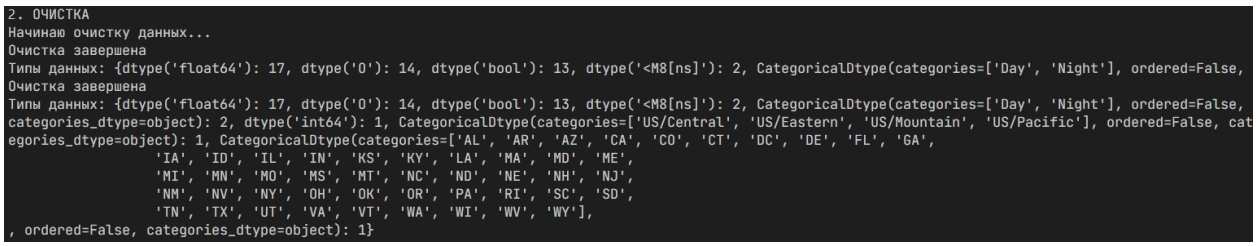

Рисунок 2 – Процесс очистки (преобразования) данных.

2.3. Load (Загрузка)

Таблица 4 – Стратегия хранения.

Формат	**Назначение**	**Объем**	**Преимущества**
Parquet	Основное хранение	7.7M записей	Сжатие, быстрая загрузка
SQLite	Тестирование	1,000 записей	Легкий доступ, SQL-запросы

3. Качество данных

Таблица 5 – Метрики качества

Метрика	**Значение**	**Оценка**	**Критерий**
Completeness	95.4%	Отлично	>70% - приемлемо
Uniqueness	100%	Идеально	>95% - отлично
Outlier Ratio	2.01%	Норма	<10% - норма

Completeness (Полнота):

- Средняя полнота данных: 95.4%
- Колонок с полнотой >90%: 47
- Колонок с полнотой <50%: 0

Uniqueness (Уникальность)

- Уникальных строк: 7,728,394
- Всего строк: 7,728,394
- Коэффициент уникальности: 100.00%

Дубликатов по ID: 0 (0.0%)

Уникальных значений по ключевым колонкам:

- Severity: 4 уникальных значений (0.0%)
- State: 49 уникальных значений (0.0%)
- City: 13,678 уникальных значений (0.2%)

- Weather_Condition: 144 уникальных значений (0.0%)
- Start_Hour: 24 уникальных значений (0.0%)

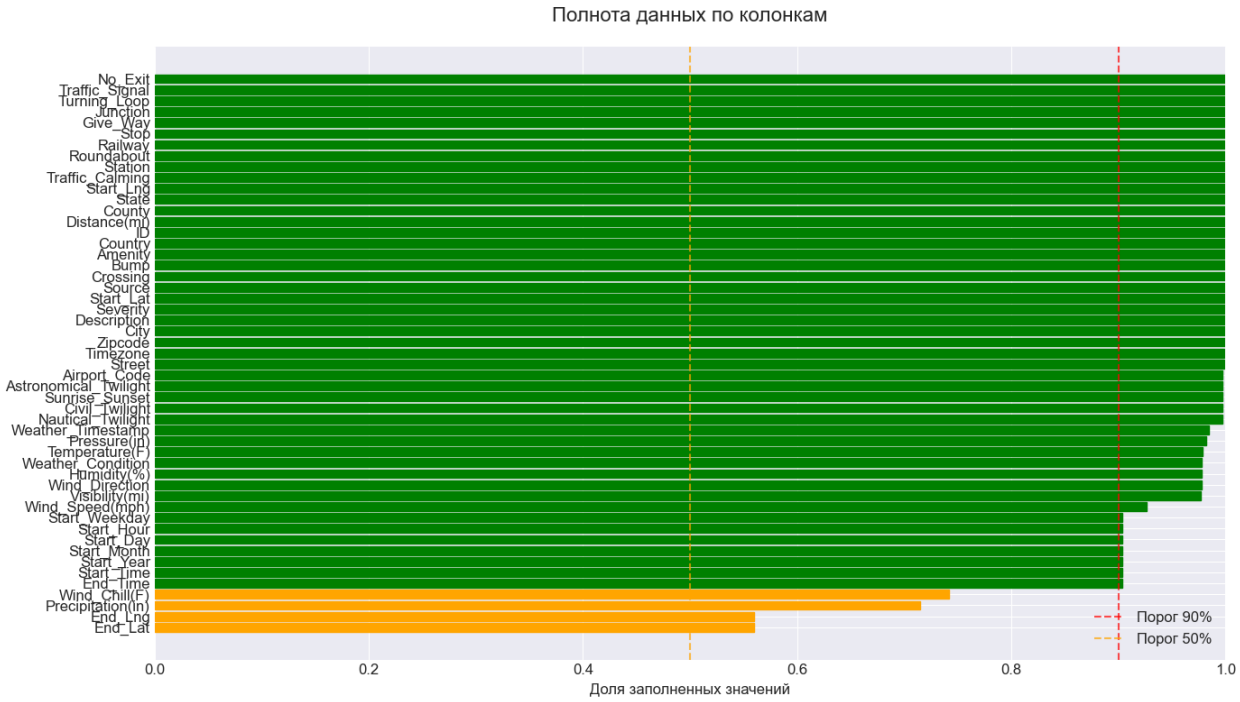

Рисунок 3 – Полнота данных по колонкам

Outlier Ratio (Выбросы):

Колонка	Выбросы	Всего значений	Доля
Distance(mi)	963,606	7,728,394	12.5%
Temperature(F)	50,515	7,564,541	0.7%
Wind_Chill(F)	43,869	5,729,375	0.8%
Итого	1,057,990	52,684,612	2.01%

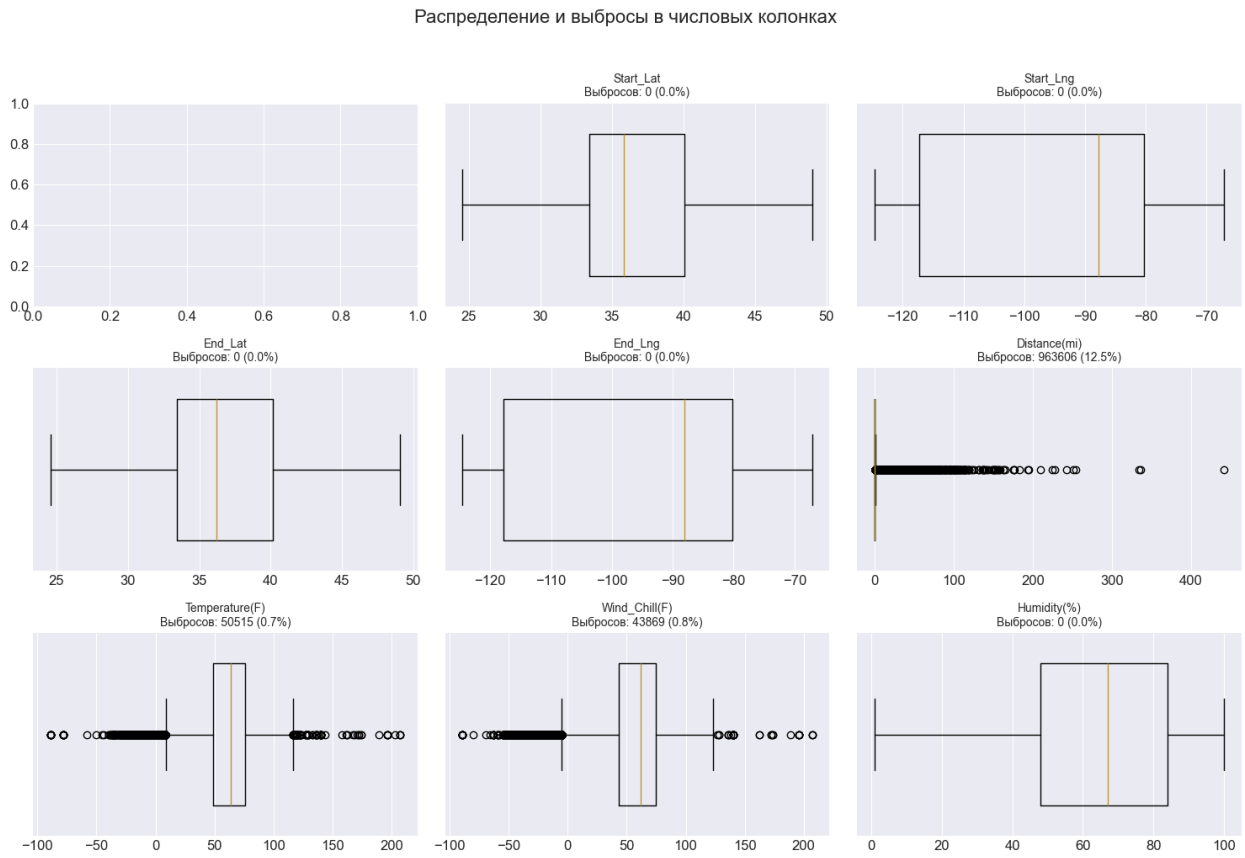

Рисунок 4 – Распределение выбросов

	Колонка	Выбросы	Всего значений	Доля выбросов
4	Distance(mi)	963606	7728394	0.124684
6	Wind_Chill(F)	43869	5729375	0.007657
5	Temperature(F)	50515	7564541	0.006678
0	Start_Lat	0	7728394	0.000000
3	End_Lng	0	4325632	0.000000
2	End_Lat	0	4325632	0.000000
1	Start_Lng	0	7728394	0.000000
7	Humidity(%)	0	7554250	0.000000

4. Анализ временных закономерностей

Изучим распределение ДТП по времени: часы, дни, недели и месяцы:

- Пиковый час аварийности: 7:00 (546,789 ДТП)
- Самый аварийный день: Пт (1,237,229 ДТП)
- Самый аварийный месяц: Дек (758,783 ДТП)

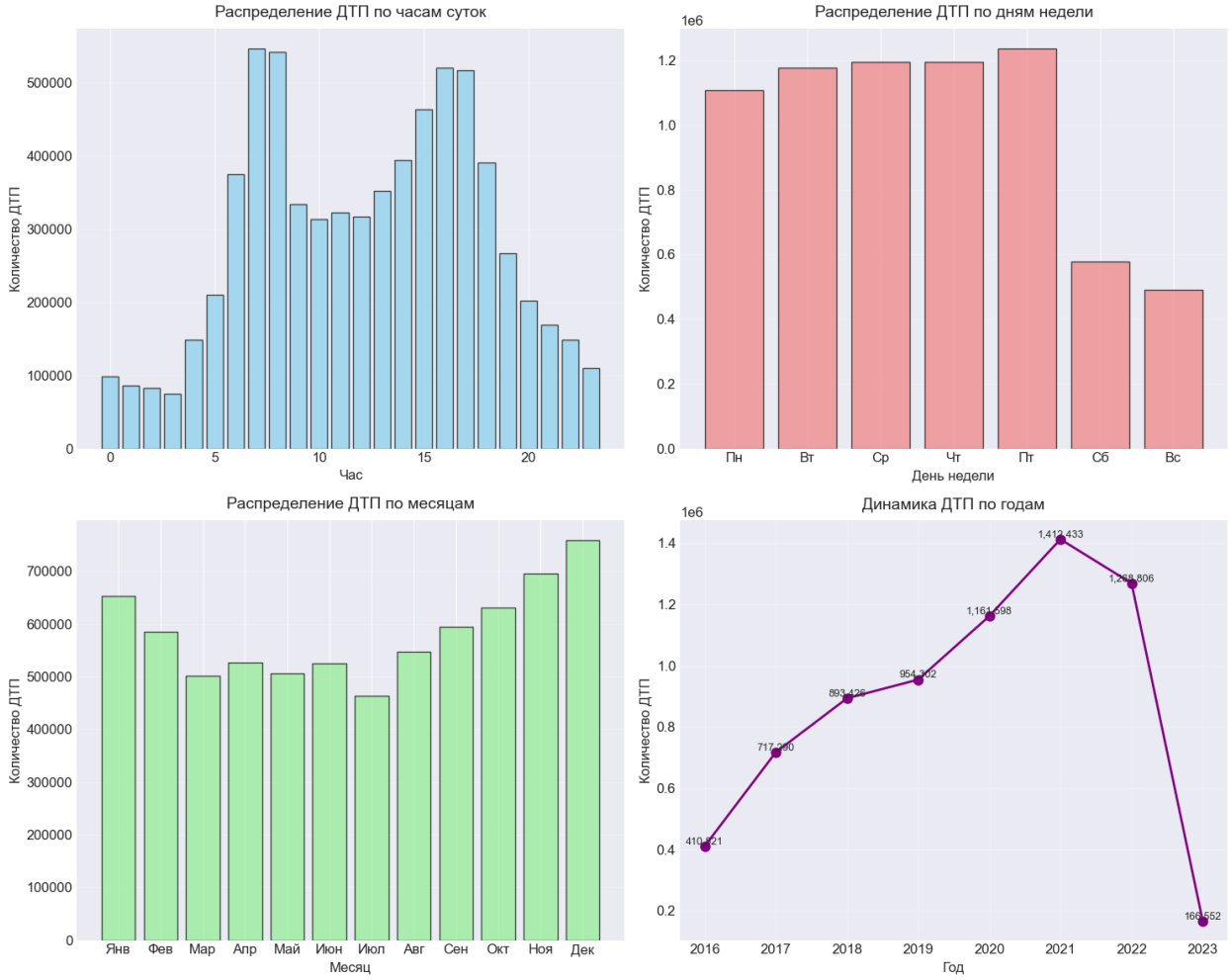

Рисунок 5 – Набор временных зависимостей.

5. Анализ тяжести аварий

Изучим распределение аварий по уровню тяжести и как тяжесть связана с другими факторами:

- Средняя тяжесть: 2.21
- Медианная тяжесть: 2
- Легкие аварии (уровень 1-2): 6,224,347 (80.5%)
- Тяжелые аварии (уровень 3-4): 1,504,047 (19.5%)

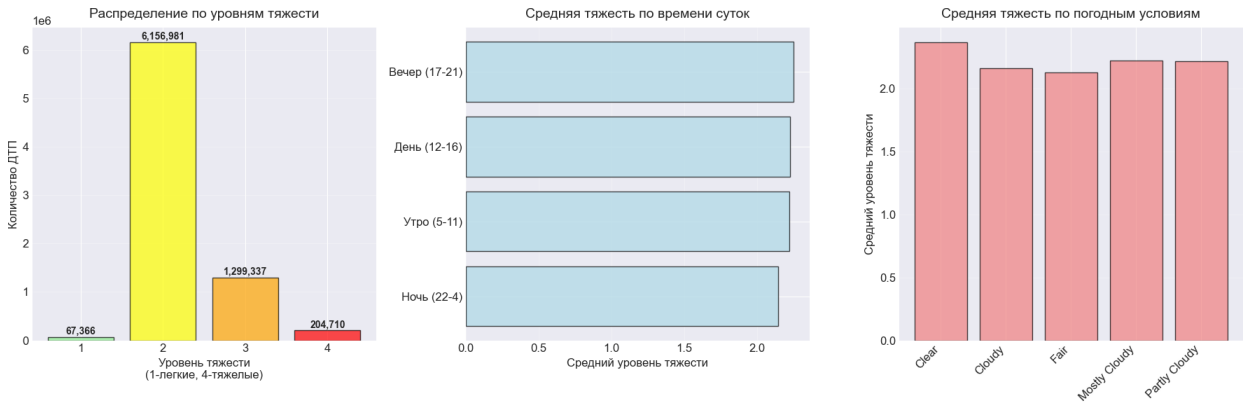

Рисунок 6 – Зависимости тяжести аварии от различных факторов.

6. Географический анализ

Проанализируем распределение ДТП по штатам и городам:

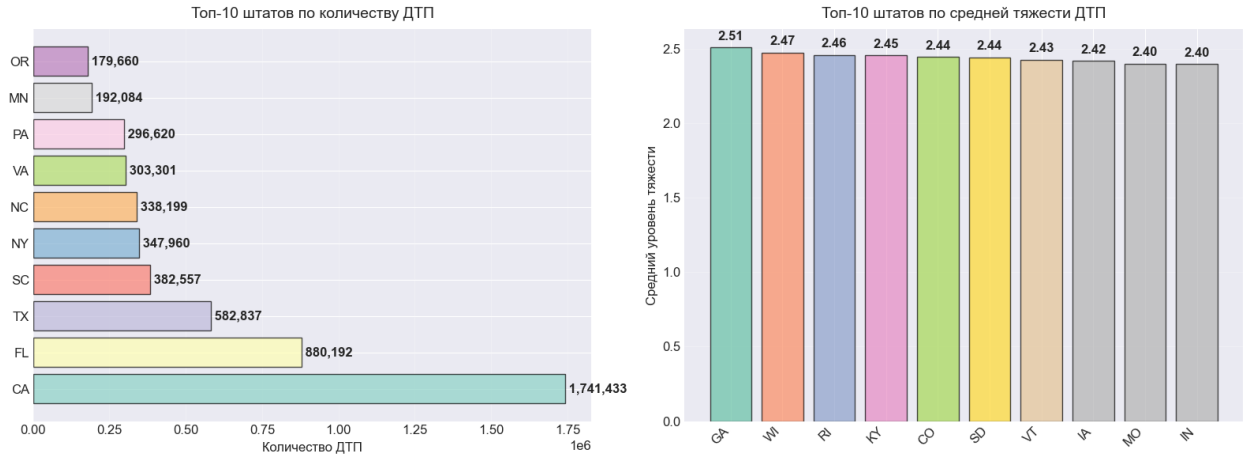

Рисунок 7 – Зависимость количества ДТП от штата.

Географическая статистика:

- Всего штатов: 49
- Самый аварийный штат: CA (1,741,433 ДТП)
- Штат с самыми тяжелыми авариями: GA (тяжесть: 2.51)
- Топ-5 городов по ДТП: Miami, Houston, Los Angeles, Charlotte, Dallas

7. Анализ погодных условий

Исследуем влияние погодных условий на частоту и тяжесть ДТП:

- Самые частые условия: Fair (2,560,802 ДТП)

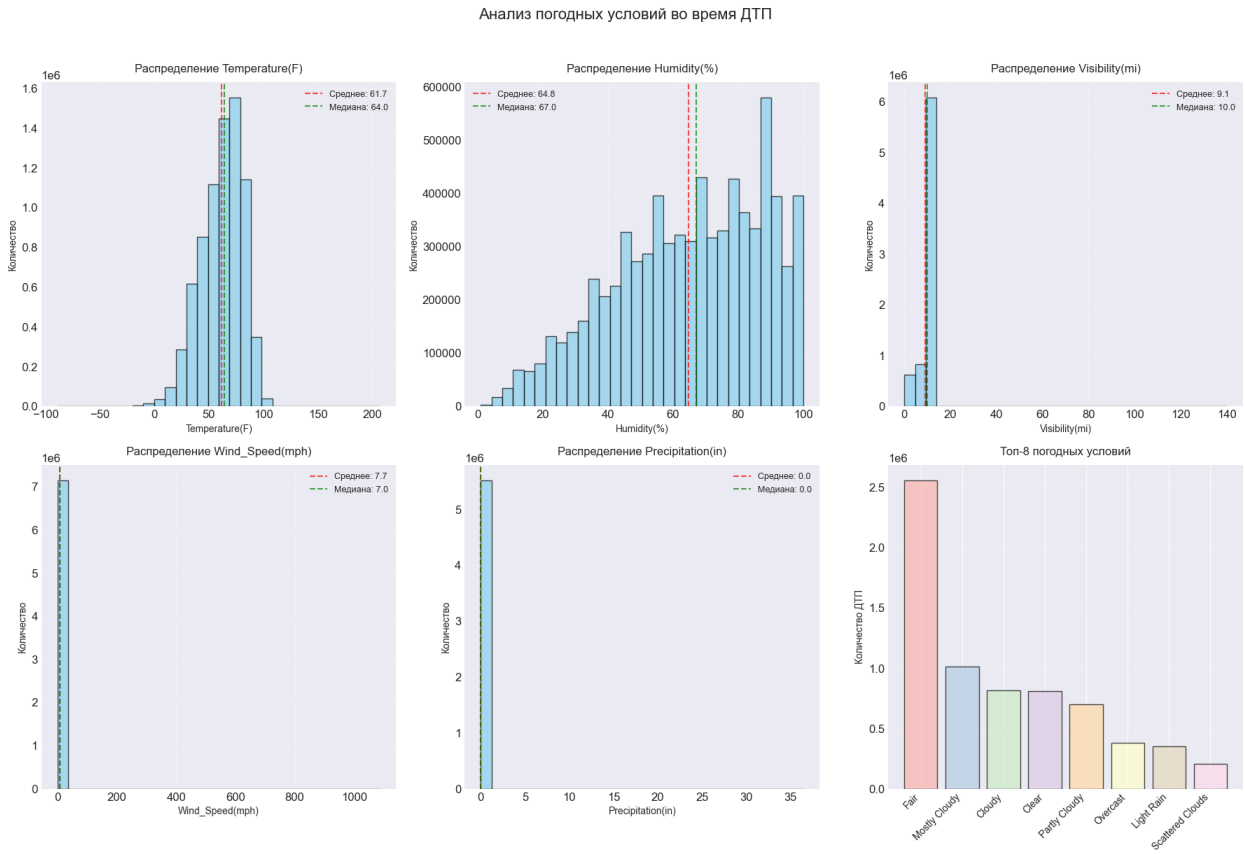

Рисунок 8 – Зависимость тяжести аварий в зависимости от погодных условий.

Корреляция тяжести с погодными условиями:

- Temperature(F): -0.020 (слабая)
- Humidity(%): 0.022 (слабая)
- Visibility(mi): -0.003 (слабая)
- Wind_Speed(mph): 0.040 (слабая)
- Precipitation(in): 0.021 (слабая)

8. Итоговые выводы

КАЧЕСТВО ДАННЫХ:

1. Полнота данных: 95.4% - ХОРОШО
2. Уникальность записей: 100.0% - ОТЛИЧНО
3. Доля выбросов: 2.0% - НОРМА

КЛЮЧЕВЫЕ ИНСАЙТЫ:

- Пик аварийности: 7.0:00 (546,789 ДТП)
- Самый опасный день: Пт
- Тяжелые аварии: 19.5% от общего числа
- Самый аварийный штат: CA (1,741,433 ДТП)

РЕКОМЕНДАЦИИ ДЛЯ ПОВЫШЕНИЯ БЕЗОПАСНОСТИ:

1. Усилить патрулирование в пиковые часы и дни недели
2. Улучшить освещение дорог в вечернее и ночное время
3. Проводить профилактические мероприятия в опасных погодных условиях
4. Сконцентрировать ресурсы в самых аварийных регионах
5. Разработать систему предупреждения водителей об опасных участках