Математическая статистика и анализ данных. Отчет по зачетной работе.

1. **Вероятность положительной задержки по авиакомпаниям**

Цель: Определить вероятность положительной задержки прилета для каждой авиакомпании и визуализировать распределение этой вероятности.

Используемые методы в реализации: Мы провели группировку данных по авиакомпаниям и вычислили среднюю вероятность положительной задержки для каждой из них. Затем был построен график, отражающий распределение этой вероятности по авиакомпаниям.

График распределения вероятности положительной задержки по авиакомпаниям демонстрирует, что некоторые из них имеют более высокую вероятность задержки, чем другие.

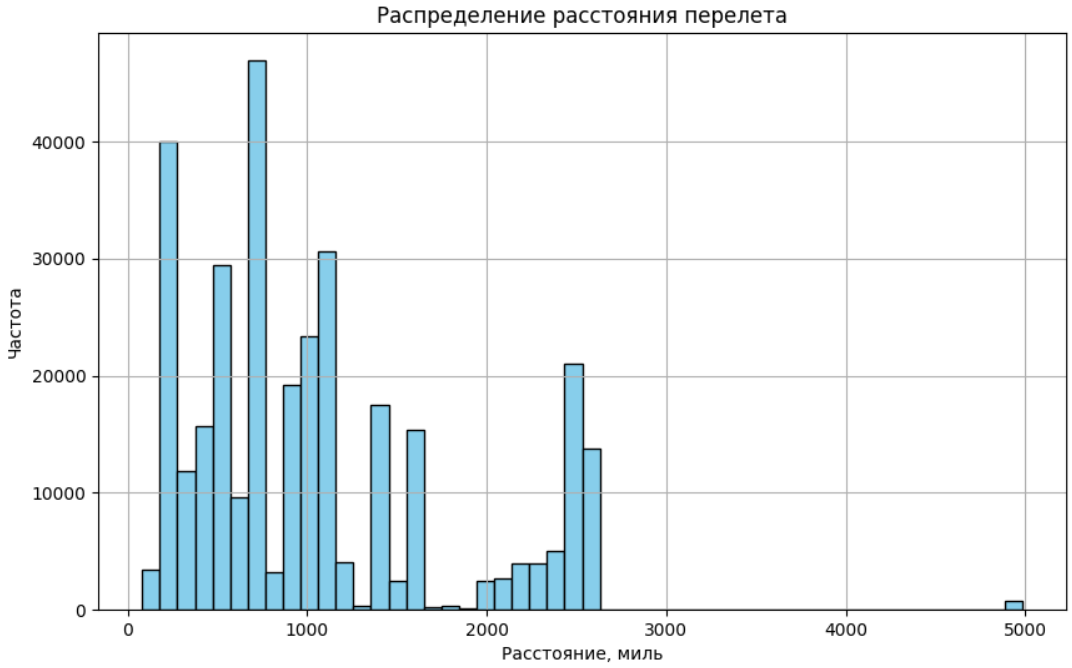
Выводыф: Эта информация может быть полезной как для пассажиров при выборе авиаперевозчика, так и для авиакомпаний с целью оптимизации своей работы и снижения задержек. Наше исследование вероятности положительной задержки прилета для каждой авиакомпании позволяет выявить различия в их операционной эффективности.

2. **Группировка перелетов по расстоянию и их задержки**

Цель: Классифицировать перелеты по расстоянию и оценить среднее время задержки вылета для коротких, средних и длинных перелетов.

Используемые методы в реализации: Группировка данных по расстоянию перелета, где каждый перелет был отнесен к одной из групп: короткие, средние или длинные перелеты. Затем для каждой группы вычислялось среднее время

задержки вылета. Анализируя полученные результаты, делались выводы о влиянии расстояния перелета на среднюю задержку вылета.



Мы построили гистограмму распределения расстояния перелетов. На основе квантилей определили три группы перелетов: короткие, средние и длинные, в качестве границ раздела мы взяли квантили уровня 0.25 и 0.5. Квантили 0.25 и 0.5 используются для определения среднего значения и медианы распределения. При таком разделении примерно половина всех перелетов попадает в категорию средних перелетов, а остальные распределяются между короткими и длинными перелетами, что и ожидается от каждого из разделов.

Квантили для определения границ групп:

0.25 509 миль

0.50 888 миль

0.75 1389 миль

Границы групп:

Короткие перелеты: до **509** миль

Средние перелеты: от **509** до **888** миль

Длинные перелеты: от **888** миль и выше

Также мы нашли все направления рейсов в случае длинного перелета(места назначения для каждого из длинных перелетов):

'IAH' 'MIA' 'BQN' 'FLL' 'MCO' 'PBI' 'TPA' 'LAX' 'SFO' 'DFW' 'LAS' 'MSP'

'RSW' 'SJU' 'PHX' 'DEN' 'SNA' 'MSY' 'SLC' 'XNA' 'SEA' 'SRQ' 'MEM' 'SAN'

'JAC' 'HNL' 'AUS' 'STT' 'EGE' 'HOU' 'LGB' 'BUR' 'MCI' 'SAT' 'PDX' 'SJC'

'OMA' 'OAK' 'SMF' 'DSM' 'PSE' 'TUL' 'OKC' 'HDN' 'BZN' 'MTJ' 'EYW' 'PSP'

'ABQ' 'STL' 'ANC'

И среднее время задержки вылета по категориям перелетов:

Категория перелета

Короткий **13.027284** мин

Средний **14.176101** мин

Длинный **11.499794** мин

Выводы:

1) Анализ классификации перелетов по расстоянию и сопутствующей задержки вылета позволяет понять, какие факторы могут влиять на операционную эффективность авиакомпаний в зависимости от длины маршрута. Мы выявили, что среднее время задержки вылета различается в зависимости от длины перелета. Эта информация может быть ценной для авиакомпаний при планировании рейсов и оптимизации процессов, а также для пассажиров, чтобы более осознанно выбирать свой маршрут и учитывать возможные задержки.
2) Самолеты в группе длинных перелетов не летят в определенные места назначения, так как каждое встречается лишь по одному разу, следовательно, то, куда они летят в малой степени статистически значимо.
3) Для длинных перелетов среднее время задержки вылета составляет около 11.5 минут. Это может свидетельствовать о том, что длинные маршруты могут быть менее подвержены задержкам по сравнению с короткими и средними маршрутами, возможно, из-за более тщательного планирования и управления рейсами на более длительные расстояния.

3. Временная динамика средней задержки вылета по месяцам

Цель: Изучить изменения средней задержки вылета в зависимости от месяца и проверить гипотезу о равенстве средних в различные месяцы.

Используемые методы в реализации: Мы использовали график средней задержки вылета по месяцам с добавлением доверительных интервалов. Для проверки гипотезы о равенстве средних в январе и феврале применили статистический t-тест для независимых выборок (Independent Samples t-test). Этот тест используется для сравнения средних значений двух групп, чтобы определить, есть ли статистически значимые различия между ними.

На графике мы отобразили среднее время задержки вылета по месяцам и добавили доверительные интервалы с уровнем доверия 95%. Статистический тест показал, что на уровне значимости 0.01 нет достаточных доказательств для того, чтобы отвергнуть гипотезу о равенстве средних в январе и феврале, но на уровне 0.05 гипотеза отвергается:

уровень значимости **0.05**: отвергается

уровень значимости **0.01**: не отвергается

Выводы: Анализ временной динамики средней задержки вылета по месяцам с использованием графика и доверительных интервалов позволяет наглядно отследить изменения во времени. Применение статистического теста для проверки гипотезы о равенстве средних значений задержки вылета в различные месяцы дает возможность оценить, действительно ли наблюдаемые различия являются статистически значимыми.

4. **Корреляция расстояния перелета и времени полета**

Цель: Исследовать связь между расстоянием перелета и временем полета и построить модель для ее оценки.

Используемые методы в реализации: Вычисление коэффициента корреляции между расстоянием и временем полета и построение точечной диаграммы для визуализации этой связи. Затем применение линейной регрессии для оценки коэффициентов и построение линии регрессии на графике.

Мы вычислили коэффициент корреляции между расстоянием и временем полета, а также построили точечную диаграмму с линейной регрессией.

Коэффициент корреляции между расстоянием и временем полета:
0.9906496472248582

Коэффициенты линейной регрессии:

slope: 0.12611932467386067

intercept: 18.466578127760357

Выводы:

1) Существует положительная корреляция между расстоянием и временем полета, что логично: чем больше расстояние, тем дольше полет. Анализ корреляции между расстоянием перелета и временем полета позволяет понять, насколько эти два параметра взаимосвязаны. На основе коэффициента корреляции и линейной регрессии можно сделать вывод о том, есть ли статистически значимая линейная зависимость между расстоянием перелета и временем полета. Полученная модель линейной регрессии может быть использована для прогнозирования времени полета на основе расстояния перелета.
2) Свободный член в уравнении линейной регрессии(intercept) - этот параметр показывает ожидаемое время в полете, когда расстояние равно нулю. Наклон линии регрессии(slope) - этот коэффициент показывает, на сколько минут увеличивается (или уменьшается) время в полете при увеличении расстояния на одну единицу(например: на одну милю).

5. Распределение задержки прилета

Цель: Определить распределение задержки прилета для рейсов, вылетевших в пределах +/-15 минут от расписания, и оценить его параметры. Построить нормированную гистограмму задержек прилета.

Предположение о распределении: Воспользуемся центральной предельной теоремой, которая утверждает, что среднее большого количества независимых и одинаково распределенных случайных величин с конечной дисперсией приближается к нормальному распределению, независимо от формы исходного распределения. В контексте задержек прилета, даже если отдельные задержки не распределены нормально, средние значения множества таких задержек могут следовать нормальному распределению.

Используемые методы в реализации: Мы построили нормированную гистограмму задержки прилета и сравнили ее с графиком плотности нормального распределения, а также оценили параметры этого распределения.

График плотности нормального распределения был нанесен поверх гистограммы для сравнения.

Оцененные параметры нормального распределения:

Среднее μ(задержка): **-8.251114540466393**

Стандартное отклонение σ: **17.907180068086404**

Выводы: Задержка прилета для данной выборки может быть описана нормальным распределением. Среднее значение задержки и стандартное отклонение позволяют оценить типичные отклонения от расписания. На основании оцененных параметров нормального распределения можно сделать предположение о характере задержек прилета в этом временном интервале. Сравнение гистограммы с графиком плотности нормального распределения позволяет оценить, что данные соответствуют этому распределению. Гистограмма

демонстрирует, что вероятность возникновения задержек прилета уменьшается с увеличением их продолжительности как справа, так и слева от центрального значения(за исключением небольшой области от нуля до среднего значения).