***Зернюк Андрей Сергеевич.***

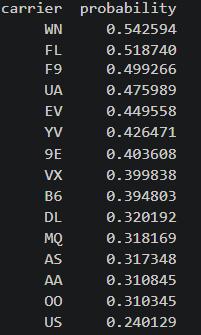
***Физический факультет 3 курс***

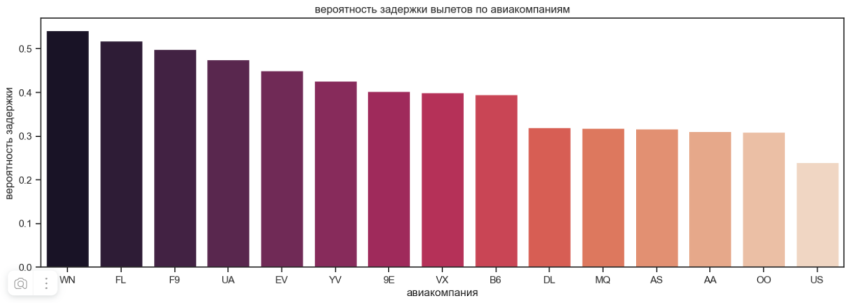
***Отчет Мфк Математическая статистика.***

**1 Задание**

По каждой авиакомпании оцените вероятность того, что задержка прилета будет положительной и нарисуйте график, показывающий распределение этой вероятности по авиакомпаниям.

Для каждой компании я нахожу количество перелетов с положительным значением задержки и делю это на общее количество перелетов.



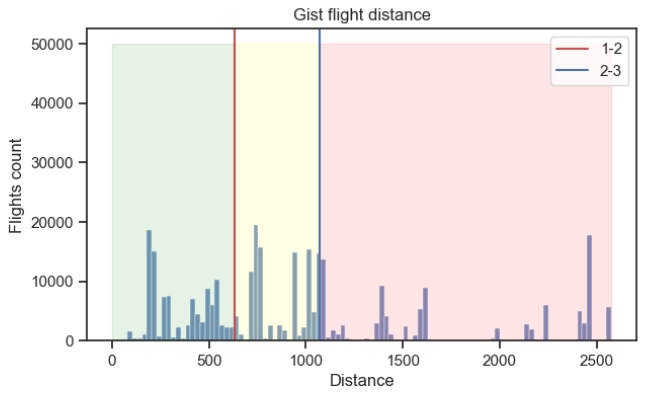


Таким образом, задержка на WN самая большая. Это исследование может быть полезно при прогнозировании времени перелета, или оценки функционирования персонала самой компании.

Я пробовал считать *отношение* суммы минут положительных задержек, делить на общую сумму минут задержек, и так как отношение может быть отрицательно, применять к результату сигмоиду. Но минус такого подхода в том, что это *отношение* лежит относительно близко к нулю и сигмоидная функция будет выдавать значение близкое к 0,5. Если бы *отношение* проектировалось на - ∞ +∞, то этот подход сработал бы.

**2 Задание**

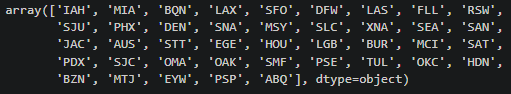
Постройте гистограмму расстояния перелета distance. Выделите на ней три группы перелетов: короткие, средние и длинные. Какие значения distance разумно выбрать в качестве границы, разделяющей короткие. средние и длинные перелеты? Куда летят самолеты в группе длинных перелетов? Найдите среднее время задержки вылета в каждой из трех выделенных групп.

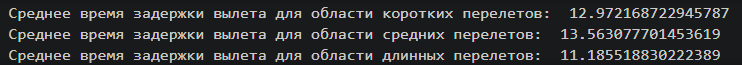


Области коротких средних и длинных перелетов я получаю, считая квантили 0.3 и 0.6

Так же удаляются выбросы (слишком длинные перелеты с помощью квантиля 0.99)

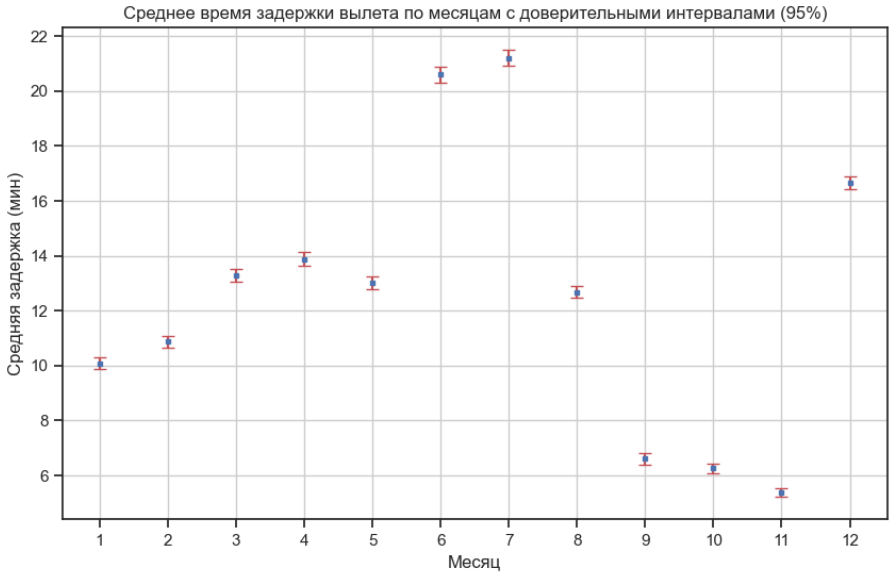
Из области длинных перелетов уникальные пункты назначения следующие





**Задание 3**

Нарисуйте график среднего времени задержки вылета по месяцам и отметьте на нем границы доверительных интервалов с уровнем доверия 0.95. С помощью подходящего статистического теста проверьте, можно ли принять гипотезу о равенстве средних в январе и феврале на уровне значимости 0.05? на уровне значимости 0.01?



Двухвыборочный t-тест используется для проверки, существуют ли статистически значимые различия между средними значениями двух независимых выборок. Сравниваем средние значения задержек вылетов в январе и феврале.

t-статистика: -2.426723604886544

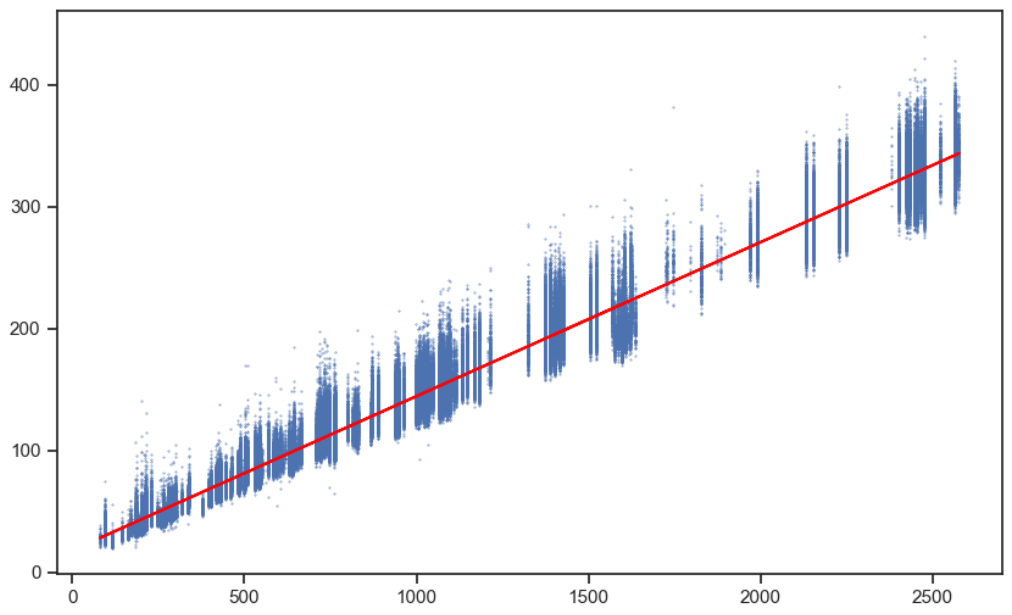
p-значение: 0.015239465878896305

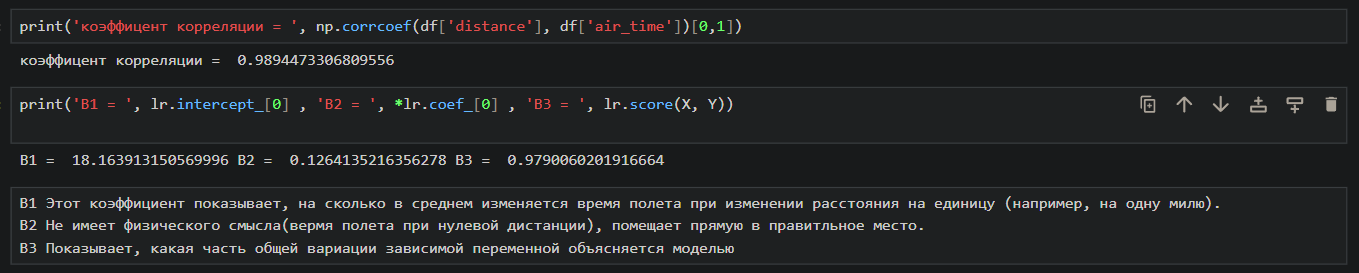
Отвергаем нулевую гипотезу на уровне значимости 0.05

Не можем отвергать нулевую гипотезу на уровне значимости 0.01

**Задание 4**

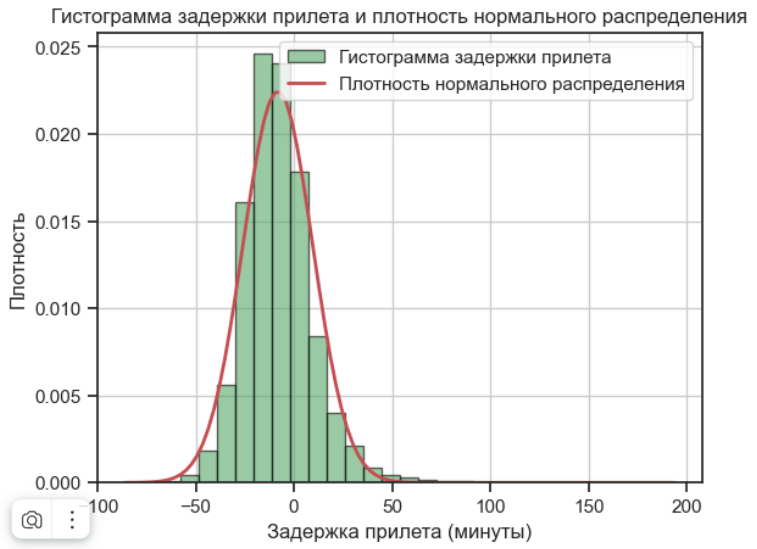
Найдите коэффициент корреляции между расстоянием distance и временем полета air\_time. Постройте точечную диаграмму в осях distance (х) и air\_time (y). Найдите коэффициенты линейной регрессии и нанесите полученную прямую на график. Как можно интерпретировать смысл коэффициентов полученной линейной модели?





**Задание 5**

Постройте нормированную гистограмму распределения задержки прилета по тем рейсам, которые вылетели в пределах +/-15 минут от времени в расписании. Сделайте предположение о том, каким распределением может описываться полученная гистограмма, оцените параметры этого распределения и нанесите график плотности на график с гистограммой. Вопрос со звездочкой (необязательный): с помощью подходящего статистического теста проверьте, можно ли принять гипотезу о выбранном распределении?



Из графика можно заметить, что данные соответствуют нормальному распределению.

**Задание 6**

Постройте модель логистической регрессии, которая будет предсказывать, что задержка прилета составит более 15 минут. Какие признаки вы будете использовать? Какие новые признаки, на основе имеющихся, можете предложить? Какова точность полученной модели (precision, recall, accuracy)? Вопрос со звездочкой (необязательный): изучите случаи, на которых модель ошибается – с чем связаны ошибки и как можно улучшить модель?

Сначала формируем таргет, как перелеты больше 15

Дополнительный признак: относительная задержка вылета и бинарный признак является ли рейс длинным(получен с применением медианы)

