

Московский Государственный Университет имени
М.В.Ломоносова
Факультет Вычислительной Математики и Кибернетики

Отчет о практической работе по МФК
"Математическая статистика и анализ
данных."

Николай Некрасов
506, ВМК МГУ

Москва, 2023

Содержание

1	Задание 1.	2
2	Задание 2.	2
3	Задание 3.	5
4	Задание 4.	5
5	Задание 5.	6
6	Задание 6.	7
7	Задание 7.	7
8	Задание 8.	8
9	Задание 9.	9
10	Задание 10.	11
11	Задание 12.	12

1 Задание 1.

Цель:

Выяснить:

- В каких колонках есть пропущенные значения
- Количество строк, в которых есть пропущенные значения
- Особенность, присущую рейсам, в которых есть пропущенные значения

Сделать:

- Удалить строки, в которых есть хотя бы одно пропущенное значение

Исследование: Для решения поставленной задачи используется библиотека pandas для языка программирования Python.

Столбцы, в которых содержатся NaN: dep_time, dep_delay, arr_time, arr_delay, tailnum, air_time.

Общее количество строк, в которых содержатся NaN: 9430.

Общей особенностью всех рейсов с NaN является тот факт, что в каждом таком рейсе отсутствует значение air_time.

2 Задание 2.

Цель:

Построить:

- Нормированные гистограммы задержек вылета и прилета в одних осях

Описать:

- Характер выбросов

Исследование: Для построения гистограмм будет использоваться библиотека matplotlib для языка программирования Python. Нахождение выбросов будет осуществляться с помощью межквартильного диапозона. На рис. 1 приведена гистограмма задержек вылетов и прилетов с исключёнными выбросами.

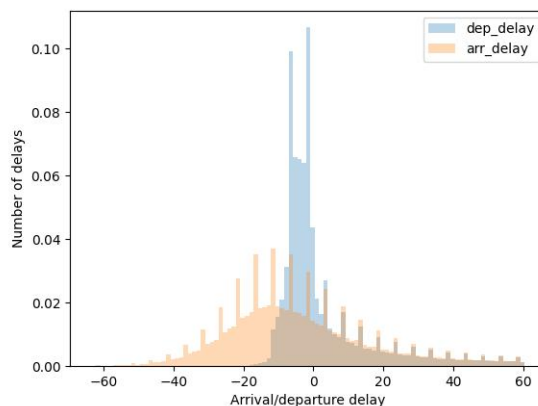


Рис. 1

На рисунках 2, 3 приведены гистограммы задержек отправления для значений меньше и больше нижней границы соответственно. На рисунках 2, 3 приведены гистограммы задержек прибытия для значений меньше и больше нижней границы соответственно.

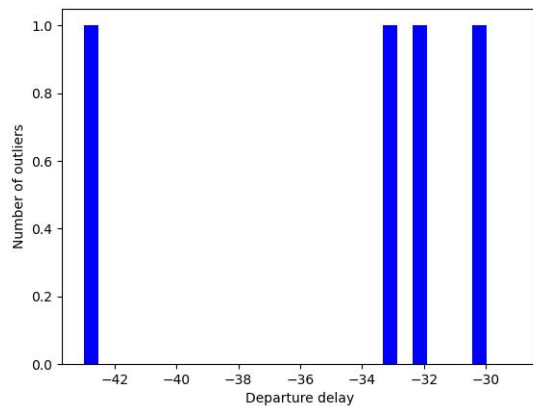


Рис. 2

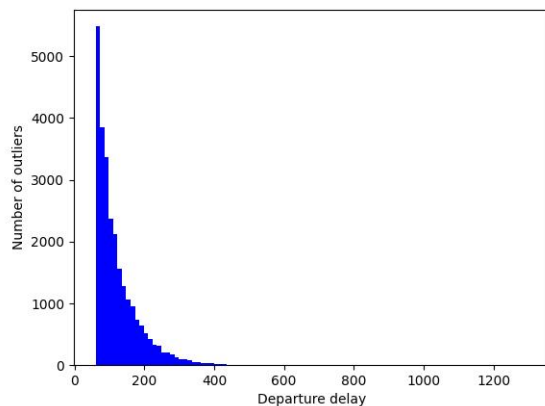


Рис. 3

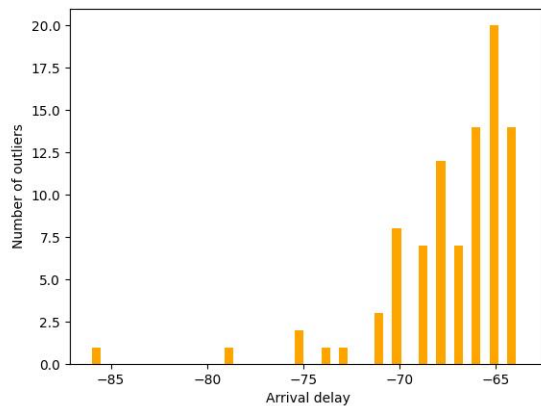


Рис. 4

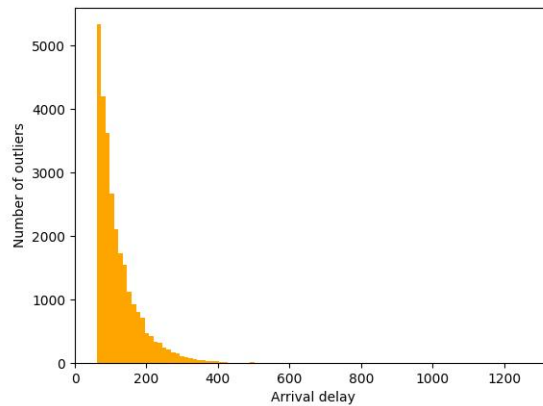


Рис. 5

Характеристика выбросов для задержек отправления, меньших нижней границы:

-30.0	1
-32.0	1
-43.0	1
-33.0	1

Характеристика выбросов для задержек отправления, больших верхней границы приведена в файле highAnomDepFile.txt по причине его большого размера.

Характеристика выбросов для задержек прибытия, меньших нижней границы:

-65.0	20
-64.0	14
-66.0	14
-68.0	12
-70.0	8
-67.0	7
-69.0	7
-71.0	3
-75.0	2
-73.0	1
-74.0	1
-86.0	1
-79.0	1

Характеристика выбросов для задержек прибытия, больших верхней границы приведена в файле highAnomArrFile.txt по причине его большого размера. Наиболее заметной особенностью распределения задержек прибытия является резкий скачкообразный рост частоты задержек через одинаковые интервалы.

3 Задание 3.

Цель:

Вычислить:

- Среднее значение
- Медиану
- Стандартное отклонение

для задержек прибытий и отправлений.

Исследование: Для решения поставленной задачи будет использоваться метод `describe()` библиотеки `matplotlib`.

Результат:

Для задержек отправления:

- Среднее значение = 12.555155706805643
- Медиану = -2
- Стандартное отклонение = 40.06568758558352

Для задержек прибытия:

- Среднее значение = 6.89537675731489
- Медиану = -5
- Стандартное отклонение = 44.63329169019399

4 Задание 4.

Цель:

Построить:

- 95%-доверительный интервал по каждой авиакомпании
- Графики доверительных интервалов для каждой компании

Отсортировать:

- Авиакомпании по средней задержке вылета

Исследование: Для построения графиков будет использоваться библиотека `matplotlib`. Доверительные интервалы будут построены с помощью формулы $\bar{X} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$, где $z_{\alpha/2} = 1.95$, s — стандартное отклонение, n — величина выборки.

Запишем компании в порядке возрастания средней задержки отправления:

US = 3.74469265291715

HA = 4.900584795321637

AS = 5.830747531734838

AA = 8.569130121764172

DL = 9.223949809056192

MQ = 10.445380836362183

UA = 12.016908379772248

OO = 12.586206896551724

VX = 12.756645817044566

B6 = 12.967547965734797

9E = 16.439574418873597

WN = 17.66165725672534

FL = 18.605984251968504

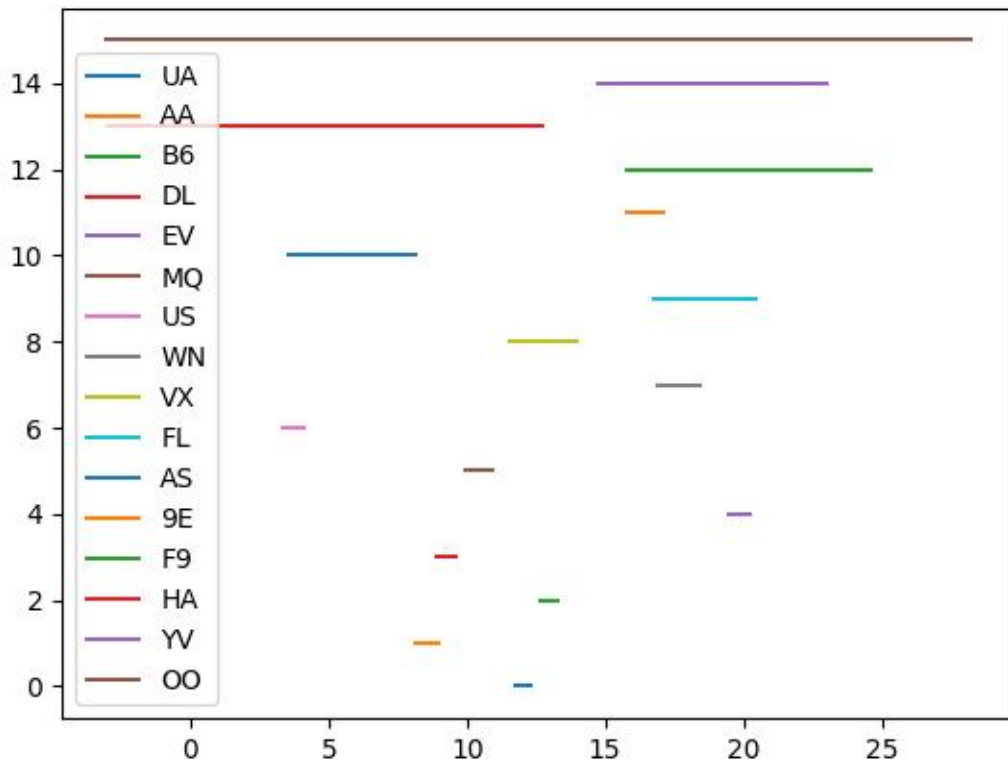
YV = 18.89889705882353

EV = 19.838929326132895

F9 = 20.201174743024964

На рис.5 построены графики этих интервалов:

Рис.5



5 Задание 5.

Цель:

Выяснить:

- Является ли различие в среднем времени задержек вылетов для авиакомпаний AL и DL статистически значимым

Исследование: Для выяснения статистической значимости будет использован Т-тест, в данном случае он применим, т.к. средние значения из соответствующих выборок распределены нормально. Т-тест реализован функцией `ttest_ind()` из библиотеки `scipy`.

Результаты: $p - value = 0.018$, тогда опровергнуть гипотезу о равенстве средних можно на уровне значимости 1.8%. Отсюда следует заключить, что различия статистически значимы.

6 Задание 6.

Цель:

Выяснить:

- Является ли различие в среднем времени задержек вылетов для аэропортов EWR, JFK и LGA статистически значимым

Исследование: Для выяснения статистической значимости будет использован Т-тест, в данном случае он применим, т.к. средние значения из соответствующих выборок распределены нормально. Т-тест реализован функцией `ttest_ind()` из библиотеки `scipy`.

Результаты: Для аэропортов JFK и LGA $p - value = 4.96 * 10^{-24}$.

Для аэропортов JFK и EWR $p - value = 3.98 * 10^{-70}$.

Для аэропортов LGA и EWR $p - value = 1.9 * 10^{-161}$.

Откуда следует, что различия статистически значимы.

7 Задание 7.

Цель:

Выяснить:

- Каким распределением описывается распределение положительных времён задержек.
- Оценить:
- Параметры данного распределения.

Исследование: Т.к. по смыслу данное распределение является дискретным, то попробуем подобрать наиболее подходящее распределение методом конечных моментов среди трёх распределений: геометрического, пуассона и биномиального. Для решения этой задачи применим функцию `logpmf` из библиотеки `scipy`.

Результат:

В результате работы функции, мы получили следующие значения для $p - value$:

Геометрическое: -594880.3176988165

Пуассона: -594880.3176988165

Биномиальное: -598015.7138514676

Откуда следует, что геометрическое распределение подходит наилучшим образом. Т.к. данное распределение является дискретным, то понятие плотности для него не определено. Гистограмма приведена на рис.6. Значение параметра $p = 0.025485840988861405$.

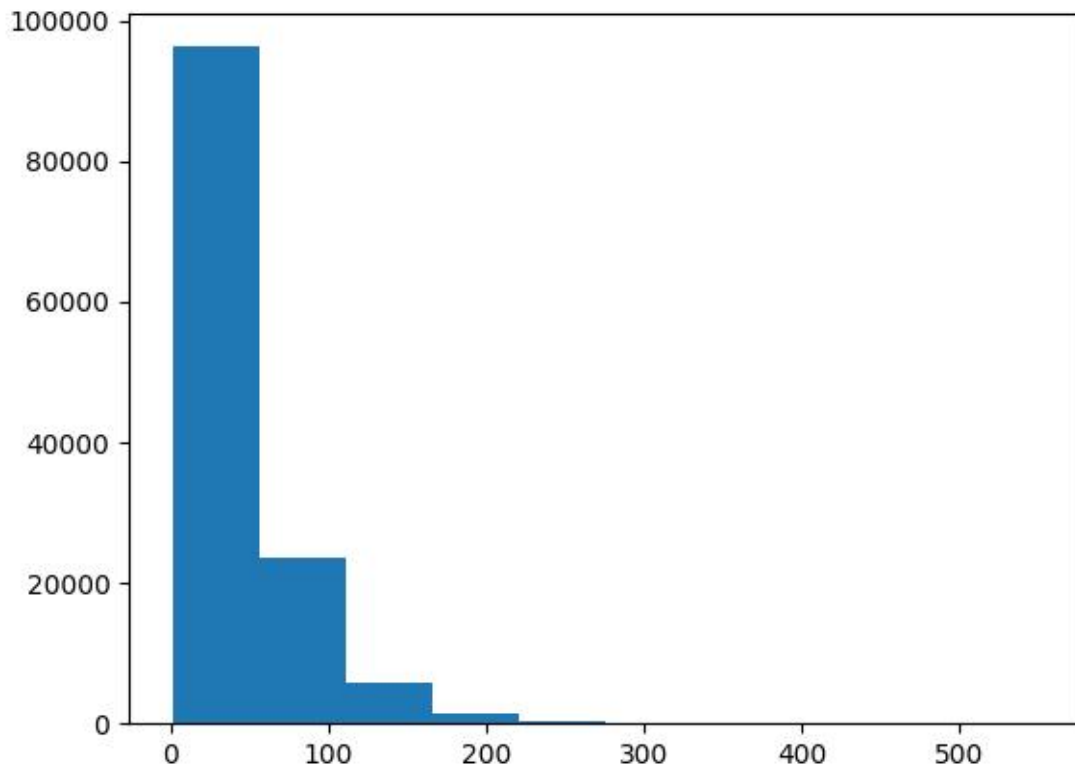


Рис. 6

8 Задание 8.

Цель:

Построить:

- Графики числа рейсов в месяц и среднего времени задержки в месяц в одних осях для тех рейсов, где величина задержки отправления положительна
- Линию регрессии на графике из предыдущего пункта

Выписать:

- Уравнение регрессии **Исследование:** Уравнение регрессии имеет вид : $y = a + bx$, где b — коэффициент корреляции, а вычисляется через b и математические ожидания данных случайных величин. Для вычисления этих коэффициентов воспользуемся функцией `corr` из библиотеки `numpy`.

Результаты: Указанный график изображен на рис. 6.

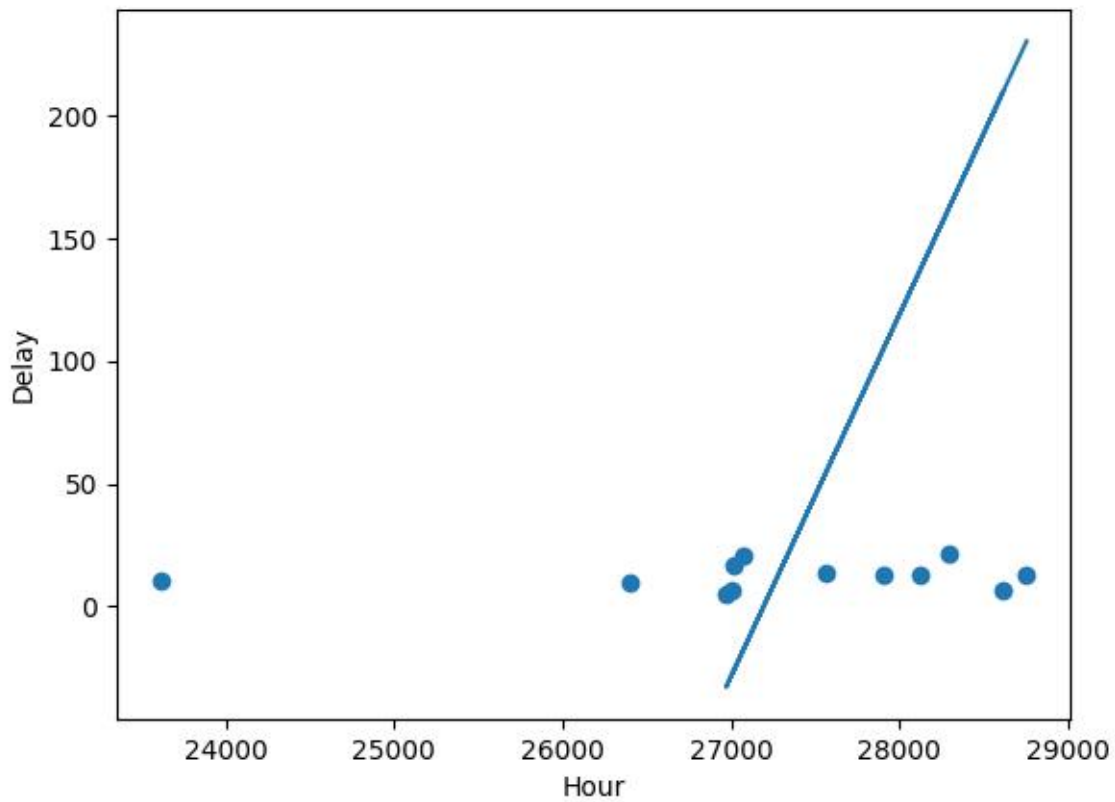


Рис. 6

Уравнение регрессии имеет вид: $y = 0.14757x - 4013.0438$

9 Задание 9.

Цель:

Построить:

- График среднего времени задержки в зависимости от часа вылета
- График доли рейсов, для которых задержка положительна, в зависимости от часа

Результат:

График среднего времени задержки в зависимости от часа вылета представлен на рис. 7.

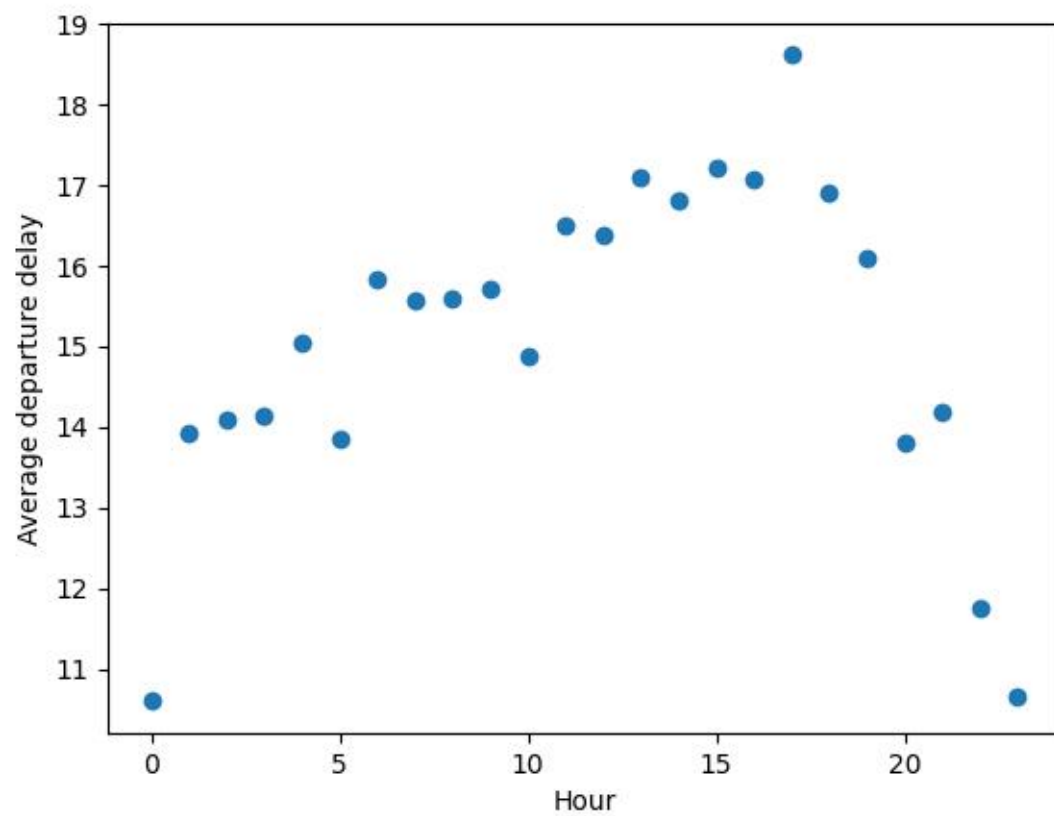


Рис. 7

График доли рейсов, для которых задержка положительна, в зависимости от часа представлен на рис. 8.

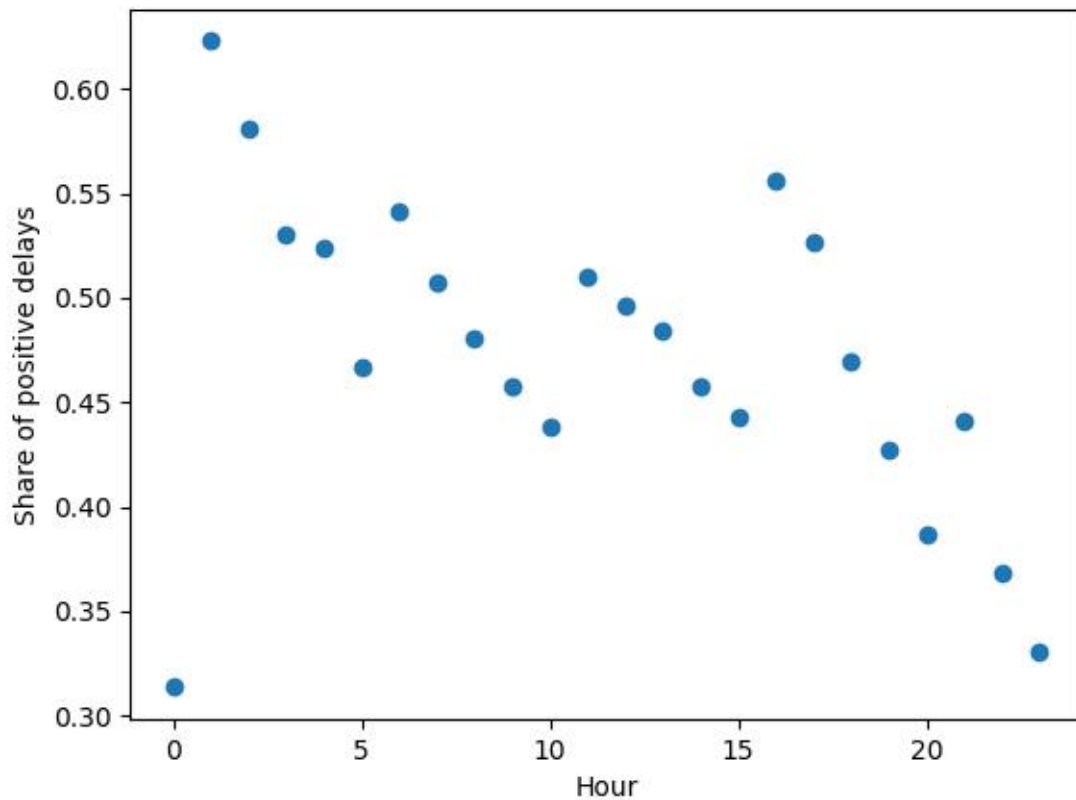


Рис. 8

Можно заметить, что наибольшая доля положительных задержек происходит в ночное, медленно снижаясь к концу дня. На графике есть два резких скачка, возможно, их можно объяснить тем, что работа в аэропортах происходит в две смены. К концу каждой из смен персонал устает, что приводит к большему числу ошибок, что, в свою очередь, увеличивает частоту задержек отправок.

10 Задание 10.

Цель:

- Выяснить:
- Какие компании являются пунктуальными
- Зависит ли предложенное разделение независимым от дальности перелета

Исследование:

Для решения данной задачи применим алгоритм кластеризации на 2 кластера по методу квадратичной ошибки.

Результат:

В результате применения алгоритма получены следующие кластеры:

Первый кластер:

AA

DL

MQ

US

AS

HA

Второй кластер:

UA

B6

EV

WN

VX

FL

9E

F9

YV

OO

Если отфильтровать входные данные, взяв дистанции меньше или больше среднего значения, то мы получим точно такие же кластеры. Поэтому пунктуальность компании не зависит от дальности полёта.

11 Задание 12.

Цель:

- Выяснить:

- Является ли различие в задержках прибытия и отправления у борта с номером 'N14228' и другими бортами статистически значимым.

Исследование: Для ответа на данный вопрос применим Т-тест. Он применим, т.к. средние значения в этих выборках распределены нормально.

Результат:

Сравнение данного с борта с другими бортами показывает, что различия статистически значимы, т.е. величины задержек прибытия и отправления зависят от номера борта. Аналогичный результат можно получить для других бортов. Значения p – *value* содержатся в файлах tailnumsDep.txt и tailnumsArr.txt.