Зачет по МФК Математическая статистика и анализ данных

Общие правила

По этой ссылке выложен датасет, содержащий информацию о совершенных внутренних авиарейсах из аэропортов Нью-Йорка в 2013 году. Ваша задача – исследовать статистику задержек авиарейсов на основе предложенных ниже вопросов. Для выполнения задания используйте любой привычный вам язык программирования или программу. Обратите внимание, что время задержки рейса может быть отрицательным и это тоже называется задержкой, если в задании нет других указаний.

Результат исследования должен быть представлен в виде структурированного отчета. В отчете по каждому вопросу нужно следовать единой логике и сначала описать, что исследуем (на какой вопрос отвечаем), затем само исследование (описать методы и модели, привести графики, таблицы, значения), и, наконец, какой вывод делаем. Все графики должны быть читаемы, снабжены необходимыми подписями и легендой. Используйте разные типы графиков для разнообразия отчета. Размерные величины всюду должны быть указаны вместе с единицами измерения.

В качестве отчета может быть представлен notebook (файл .ipynb), содержащий код исследования, графики и текстовое описание. Notebook должен быть выложен в репозиторий (gist, github, gitlab, …) или google colab. Это предпочтительный вариант. Другой вариант – оформить отчет в виде файла pdf и приложить исходный код (например, файлы .c, .m, таблицу excel с расчетами и т.п.). Ссылки на файлы необходимо отправить через гугл-форму https://forms.gle/TwbXyi11txftVT12A. Обратите внимание, что если к файлам нет доступа на просмотр, проверка работы проходить не будет.

Проверка работ будет проходить в даты зачетов (17, 24 и 31 мая). Проверяются работы, присланные строго до дня очередного зачета. Результаты в течение дня зачета будут публиковаться в таблице по этой ссылке. По расписанию зачетов можно прийти проставить полученный зачет в зачетку. Можно сдавать досрочно и договариваться о досрочной проверке. При проверке работ в первую очередь будет оцениваться грамотность работы с данными. Зачет не является соревнованием в точности моделей, но и не ограничивает вас в вопросе совершенствования моделей. Необходимый минимум для зачета – 10 из 12 заданий.

Описание датасета

Датасет представлен файлом csv, в котором 336 776 строк и 13 столбцов, содержащих
year, month, day: date of departure;
dep_time, arr_time: actual departure and arrival times (format HHMM or HMM), local timezone;
dep_delay, arr_delay: departure and arrival delays, in minutes. Negative times represent early departures/arrivals;
carrier: two letter carrier abbreviation;
flight: flight number;

tailnum: plane tail number;
origin, **dest:** origin and destination airports;
distance: distance between origin and destination airports.

(Справочно) Аббревиатуры аэропортов Нью-Йорка:
JFK – John F. Kennedy International Airport;
LGA – LaGuardia Airport;
EWR – Newark International Airport.

Задания для исследования

1. В каких колонках есть пропущенные значения? Сколько строк, в которых есть хотя бы одно пропущенное значение? Есть ли какая-то особенность в тех рейсах, в которых есть пропущенные значения? Удалите строки, в которых есть хотя бы одно пропущенное значение из дальнейшего анализа.

2. Постройте в одних осях нормированные гистограммы времени задержки вылета и прилета. Ограничьте диапазон построения гистограмм, чтобы избавиться от выбросов, и опишите характер выбросов (количество и значения). Есть ли другие особенности в полученных распределениях?

3. Оцените среднее значение, медиану и величину стандартного отклонения для времени задержки вылета и времени задержки прилета.

4. Отсортируйте авиакомпании по величине средней задержки вылета и приведите среднюю задержку вылета вместе с 95%-доверительным интервалом по каждой авиакомпании. Результат представьте в виде графика.

5. Значимо ли различие в среднем времени задержки вылета для авиакомпаний American Airlines (AA) и Delta Airlines (DL)? На каком уровне значимости можно отвергнуть гипотезу о равенстве средних?

6. Сравните между собой аэропорты вылета (JFK, LGA, EWR) с точки зрения статистики задержек вылетов. Являются ли различия статистически значимыми?

7. Каким распределением можно описать распределение времени задержки вылета в диапазоне, где время задержки вылета > 0? Предложите общий вид распределения и оцените его параметры. На одном рисунке изобразите гистограмму и график плотности аппроксимирующего распределения.

8. Для тех рейсов, для которых задержка вылета > 0, постройте в одних осях графики числа рейсов в месяц и среднего времени задержки в месяц. Найдите коэффициент корреляции между полученными значениями. Постройте точечную диаграмму (scatterplot), показывающую зависимость между полученными значениями (по оси X отложите число рейсов в месяц). Нанесите на график линию регрессии. Выпишите уравнение регрессии.

9. Постройте график среднего времени задержки в зависимости от часа вылета. На отдельном графике постройте долю рейсов, для которых задержка > 0, в зависимости от часа вылета. Опишите словами наблюдаемую картину.

10. Предложите способ разделить авиакомпании на пунктуальные и непунктуальные. Какие авиакомпании в какую группу попадают? Будут ли различаться группы в зависимости от дальности перелета?

11. Предложите модель, дающую прогноз средней задержки вылета по каждому аэропорту на следующий день. Какие признаки вы будете использовать, на каких данных будете обучать (настраивать) модель, на каких тестировать? Оцените точность вашей модели. В частности, сделайте прогноз на 31 декабря 2013 года и сравните с фактическим значением (исключите эту дату из обучающей и тестовой выборки).

12. Приведите еще хотя бы один интересный факт из датасета, раскрывающий или характеризующий особенности в возникновении/распределении задержек авиарейсов.

Успехов!