

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ

“КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ

імені ІГОРЯ СІКОРСЬКОГО”

Факультет прикладної математики

Кафедра програмного забезпечення комп’ютерних систем



**Лабораторна робота №6**

з дисципліни: «Технології оброблення великих даних»

на тему: «Розподілені обчислення даних з використанням Spark-кластера у середовищі R»

Виконав

студент IІІ курсу каф. ПЗКС ФПМ

групи КП-82

Мельничук Олексій Геннадійович

Перевірила

доц. каф. ПЗКС ФПМ

Олещенко Л.М.

Київ 2021

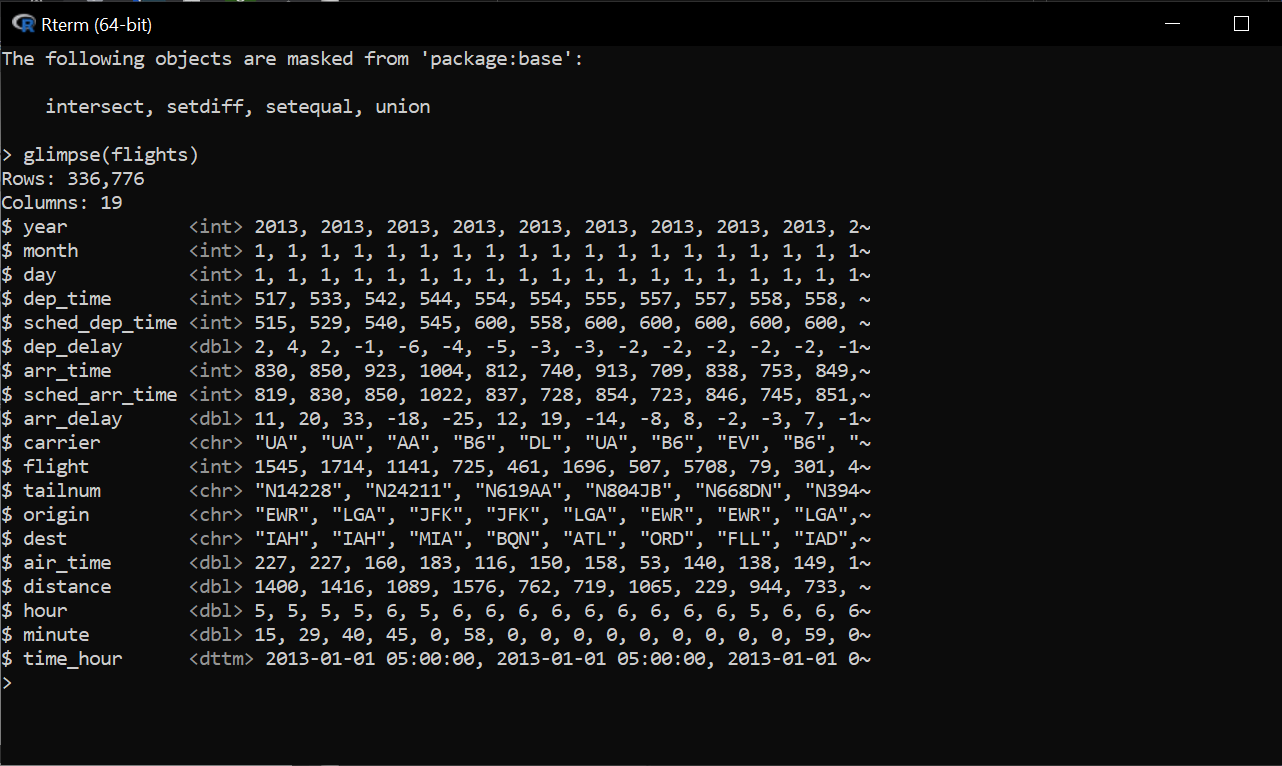
1. **Індивідуальне завдання**

**Мета**: встановити Spark на локальній машині, ознайомитись з можливостями розподілених обчислень для великих даних з використанням Spark-кластера у середовищі R.

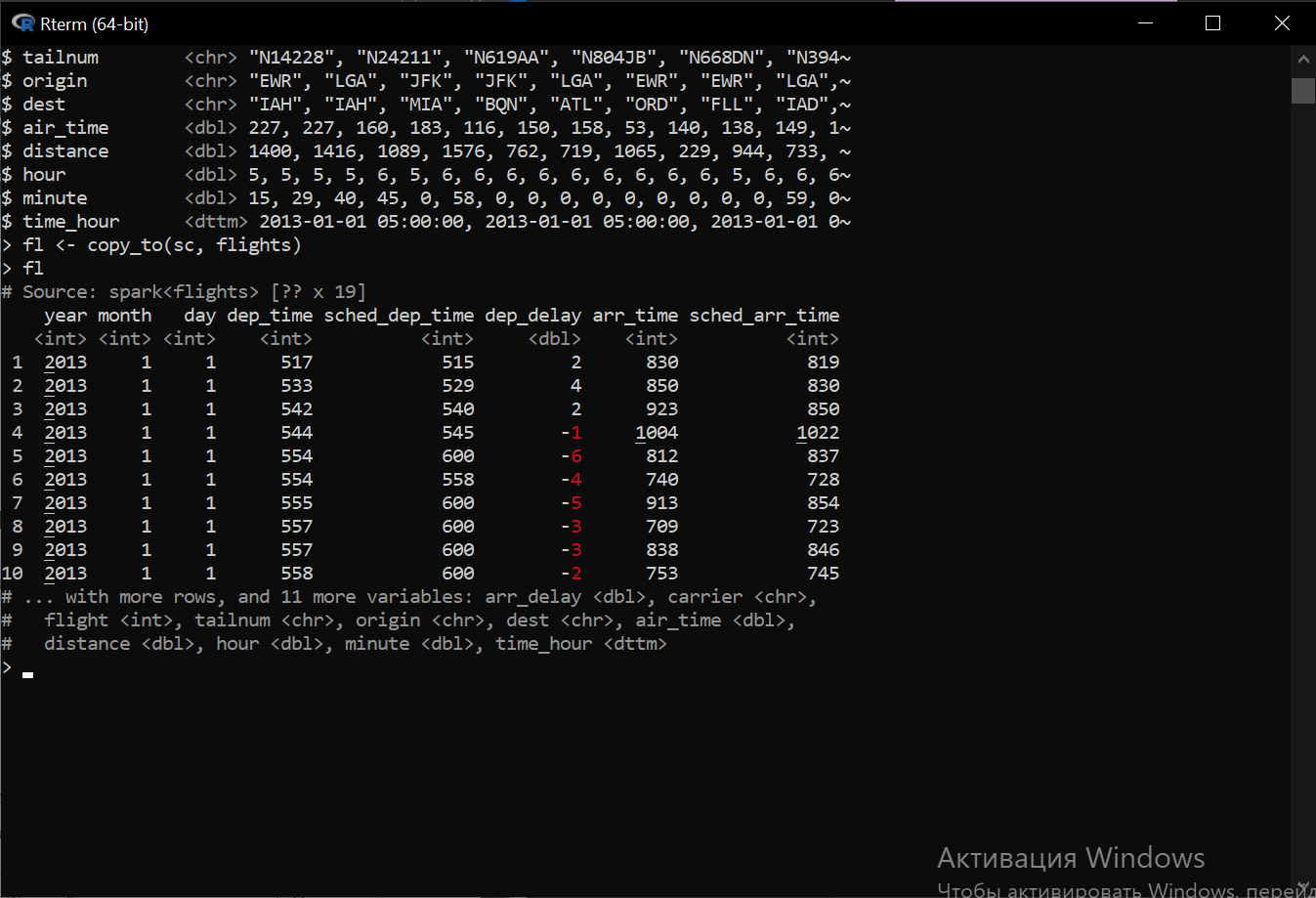
1. **Хід роботи**

**Частина 1: Встановлення та підключення до локального Spark-**

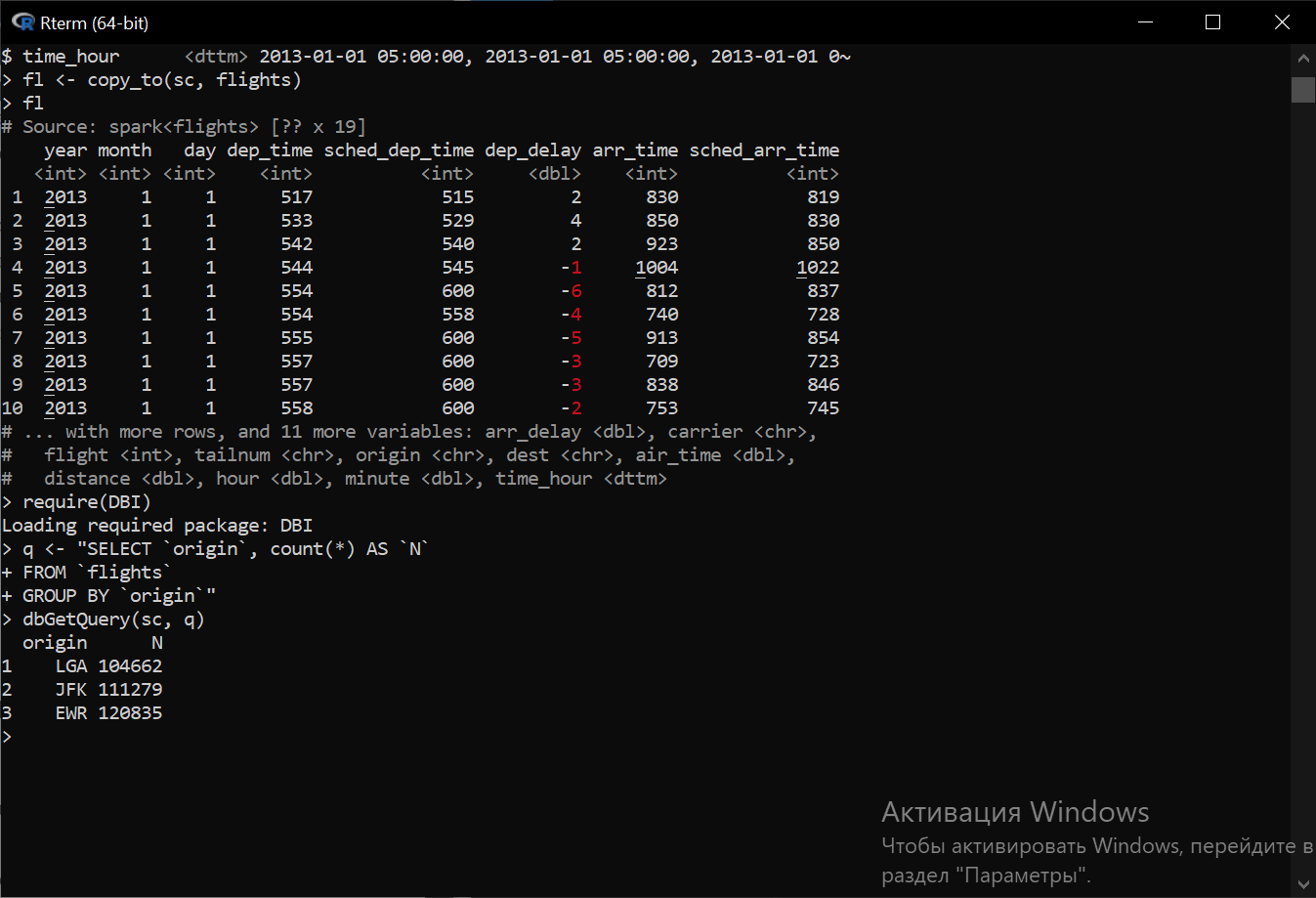
**кластеру**

Для лабораторної роботи використаємо дані одного з багатьох пакетів даних, пакета nycflights13, який містить кілька таблиць з описом 336776 авіарейсів, які стартували з аеропортів Нью-Йорка в 2013 р. Встановіть пакет nycflights13

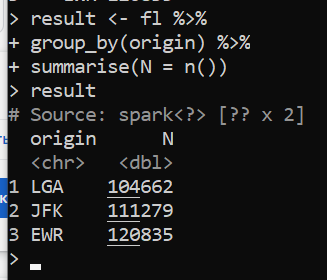
Для початку потрібно завантажити таблицю flights в локальний Spark-кластер. Для цього служить функція copy\_to (), на яку ми подаємо створений раніше об'єкт sc і копійований таблицю flights



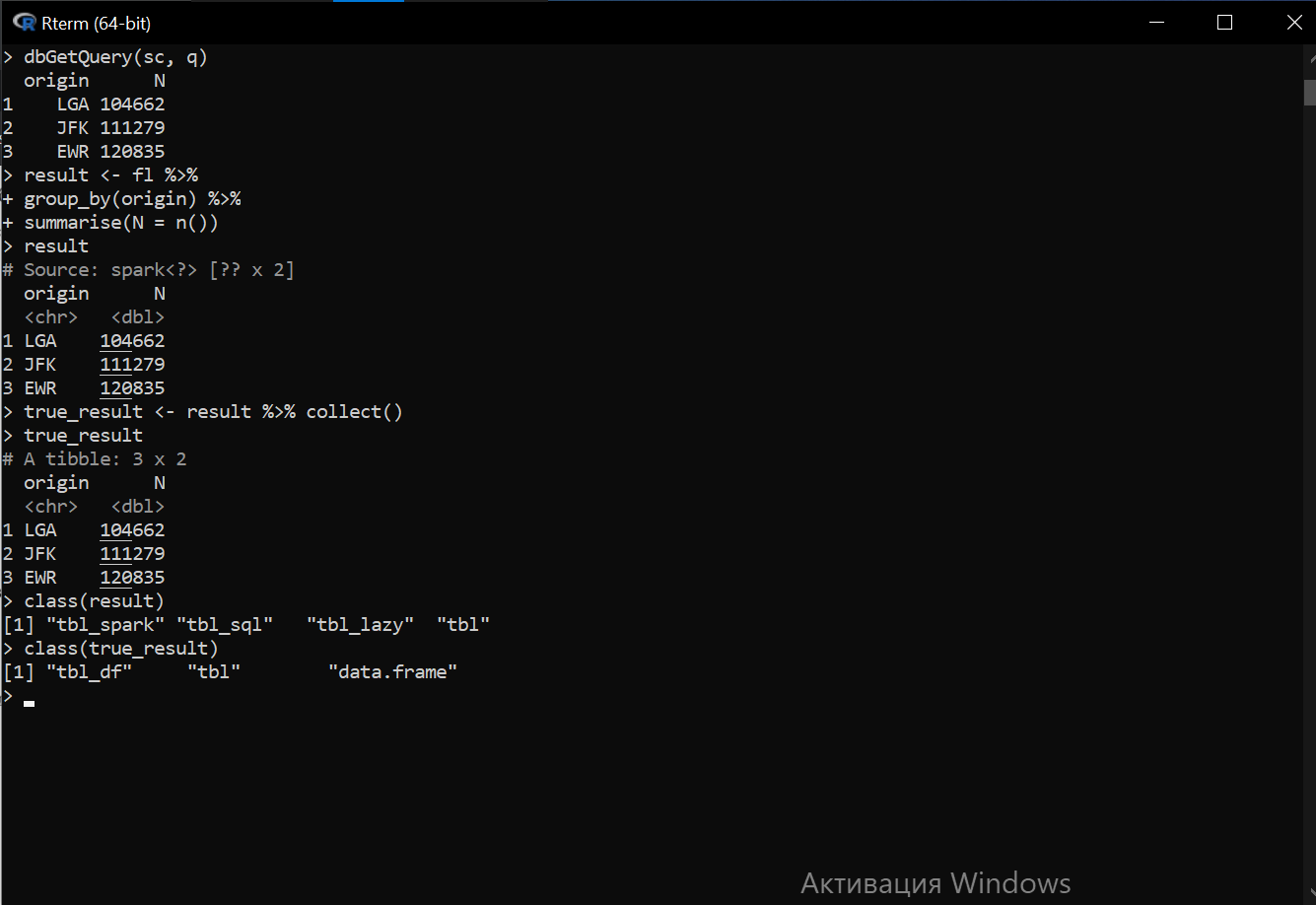
Для виконання SQL-запиту потрібно пакет DBI і входить до його складу функція dbGetQuery () . Якщо ви використовували цей пакет раніше для роботи з віддаленими базами даних, то не зустрінете нічого незвичайного: на функцію dbGetQuery () потрібно подати об'єкт sc з інформацією для підключення до Spark- кластеру, а також сам запит:



Використання SQL майже потрібно при виконанні складних запитів, які повинні бути оптимізовані для забезпечення швидкодії. Однак в більшості випадків для виконання стандартних операцій над даними використання функцій з пакета dplyr буде більш зручним. Більш того, якщо ви раніше працювали з dplyr (а сьогодні навряд чи знайдуться користувачі R, які не знають про цей пакет), то ви не побачите майже ніяких відмінностей:

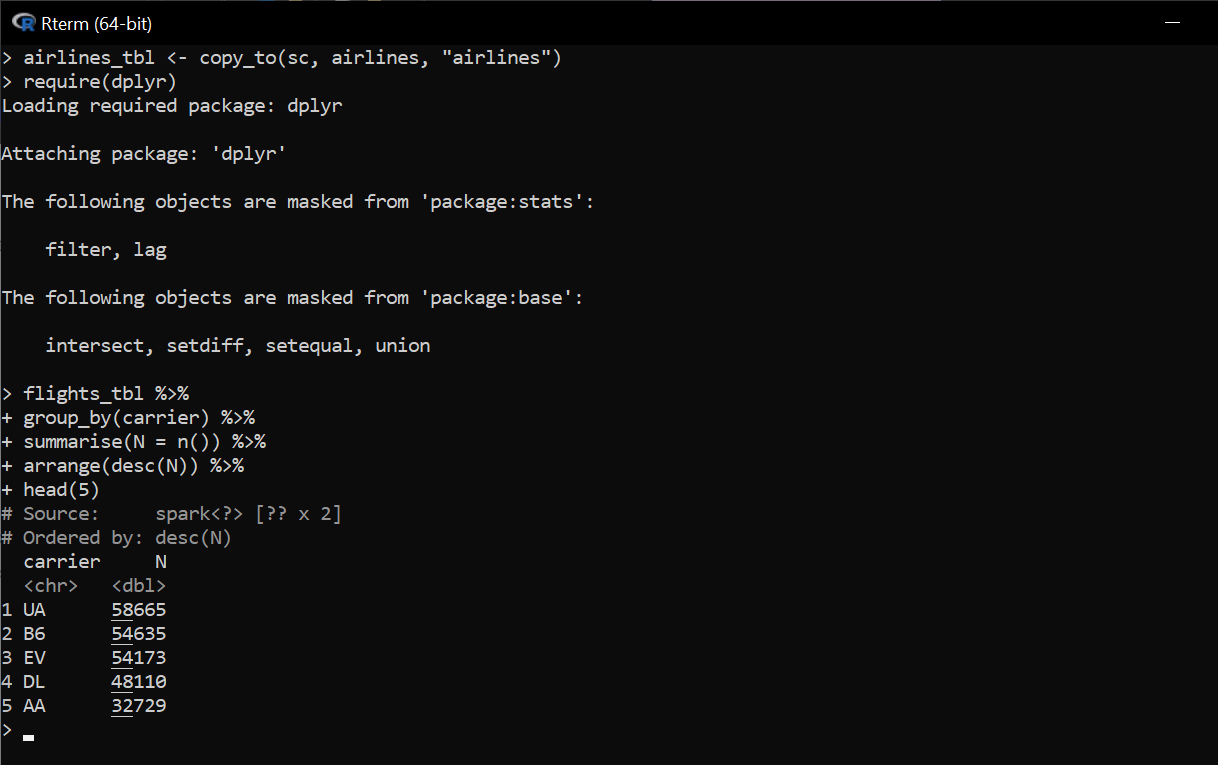


Для отримання даних з Spark в середу R слід скористатися функцією collect():



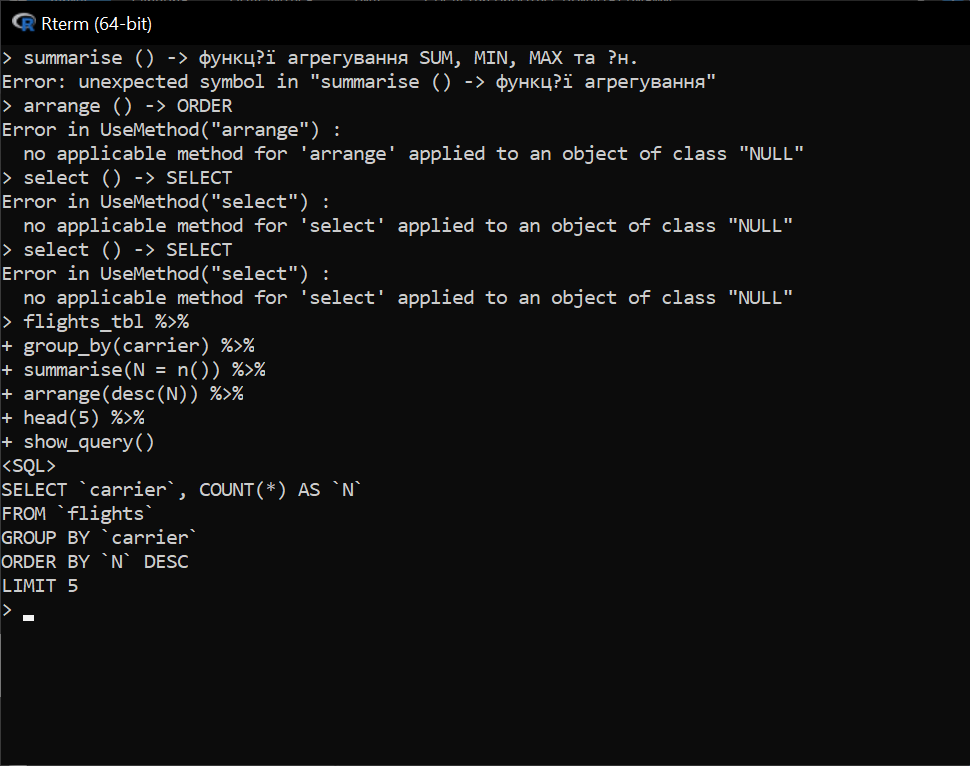
**Частина 2: Аналіз даних в Spark-кластері за допомогою пакета dplyr в R**

Підрахуйте загальну кількість польотів, виконаних кожною авіакомпанією за 2013 р., а потім виберіть 5 авіакомпаній з найбільшим числом польотів:



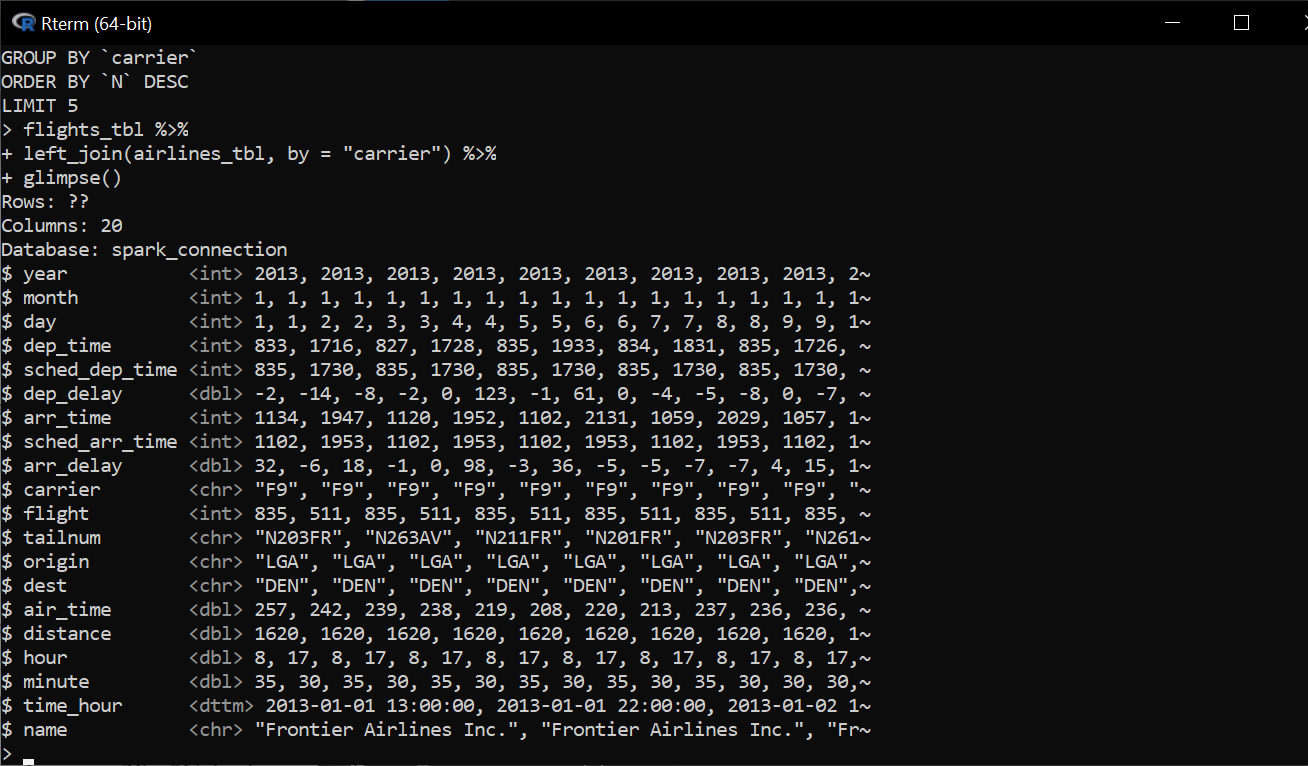
Функція show\_query () з пакету dplyr дозволяє переглянути SQL-запит, який

формується з відповідного коду R. Для наведеного вище прикладу отримуємо:



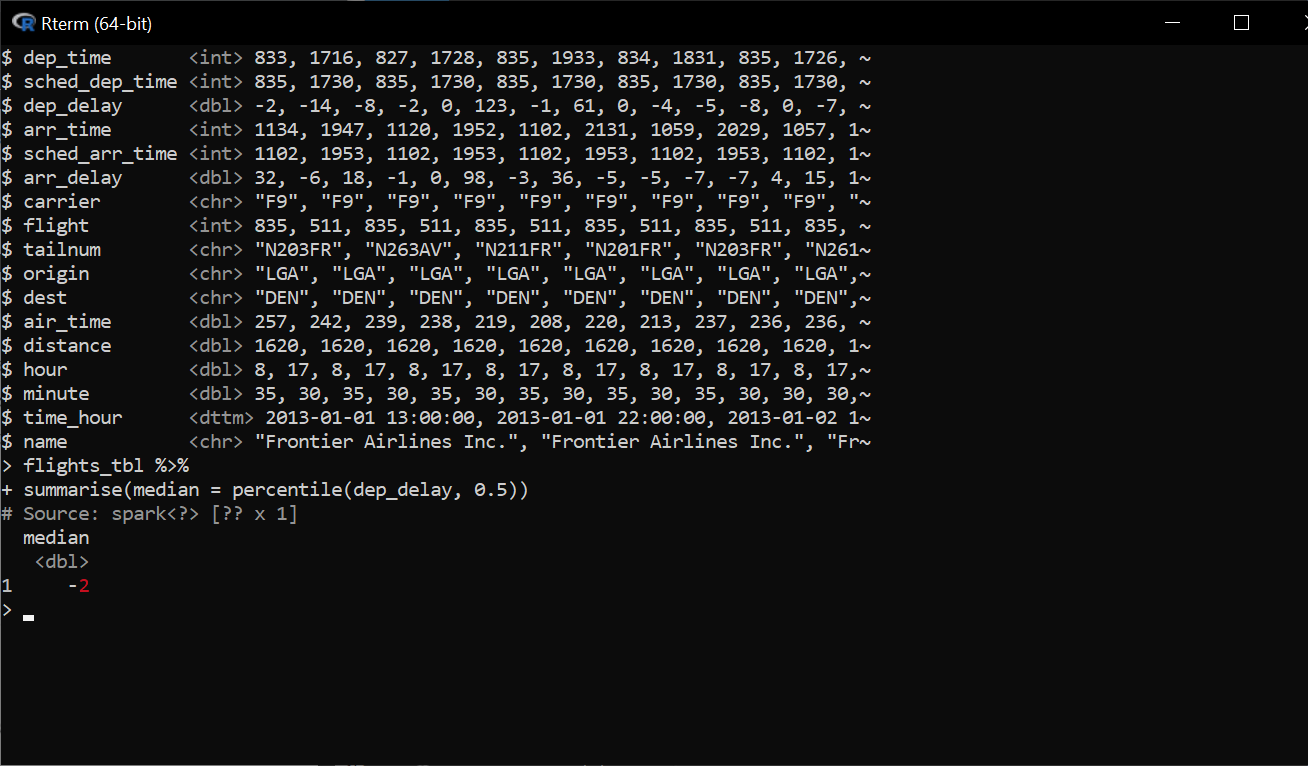
**Об'єднання таблиць**

Так, в наступному прикладі виконаний LEFT JOINтаблиці flights\_tbl з таблицею airlines\_tbl по полю carrier :

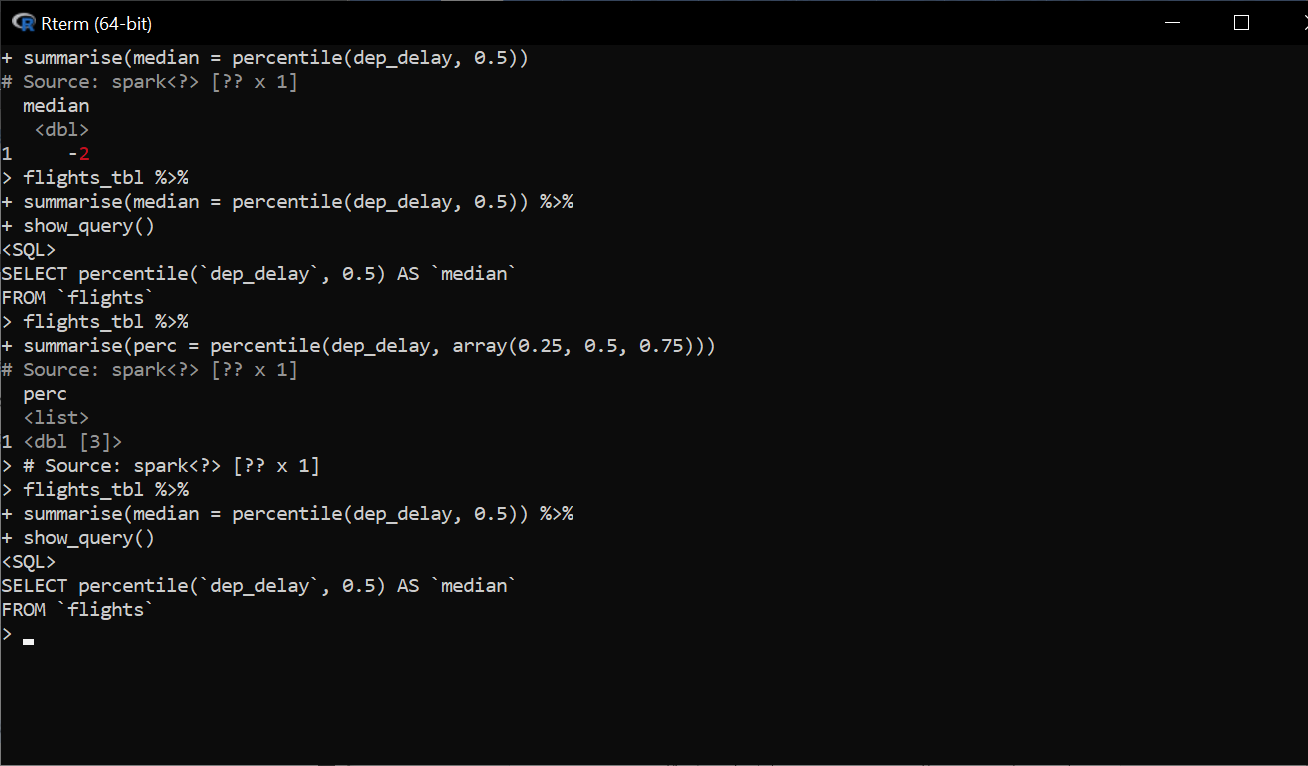


**Функції Hive Query Language**

Наприклад, для обчислення медіанного значення змінної dep\_delay (затримка рейсу, хв) з таблиці flights\_tbl ми не можемо просто скористатися базовимифункціями R median () або quantile () – це призведе до помилки. Але ми без проблем можемо застосувати Hive-функцію percentile ():



Коли при перекладі коду R на SQL dplyr зустрічає незнайому функцію, то він просто включає її в SQL-запит "як є":



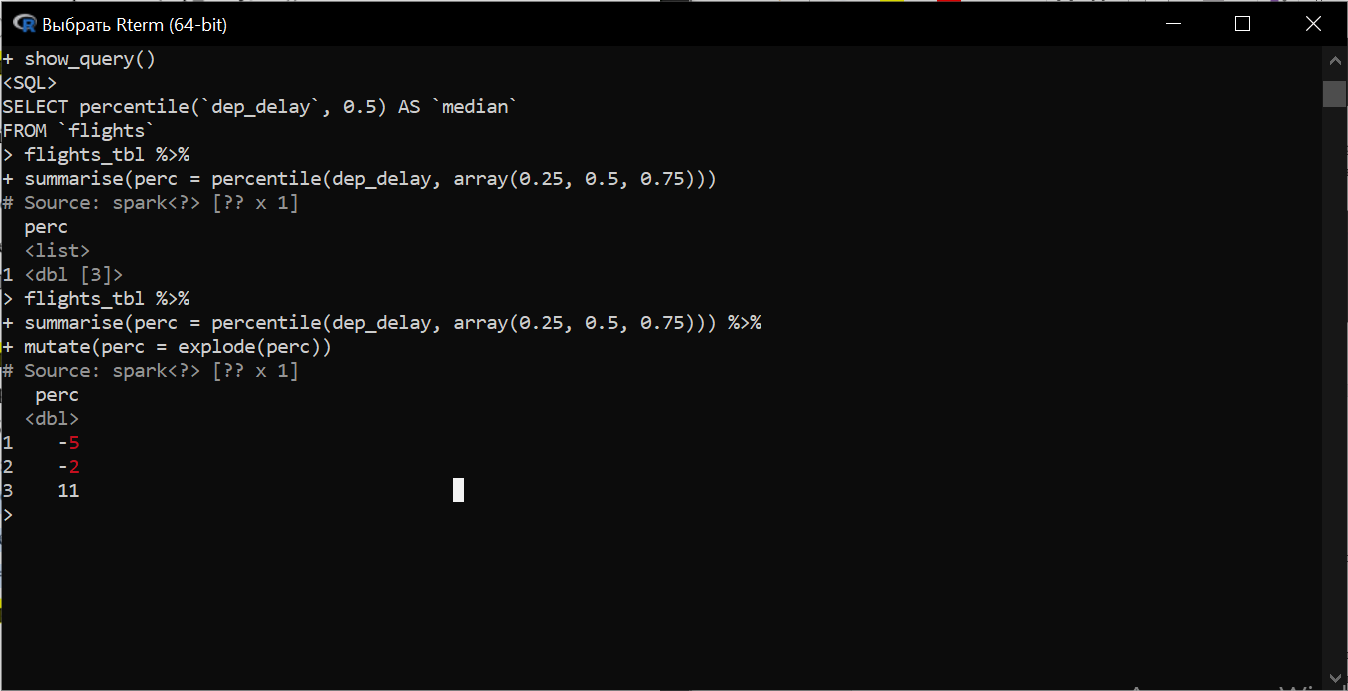
Hive-функція percentile () дозволяє одночасно обчислити кілька процентилей. Для цього на неї потрібно подати масив (array ()) з необхідними значеннями

процентилей:

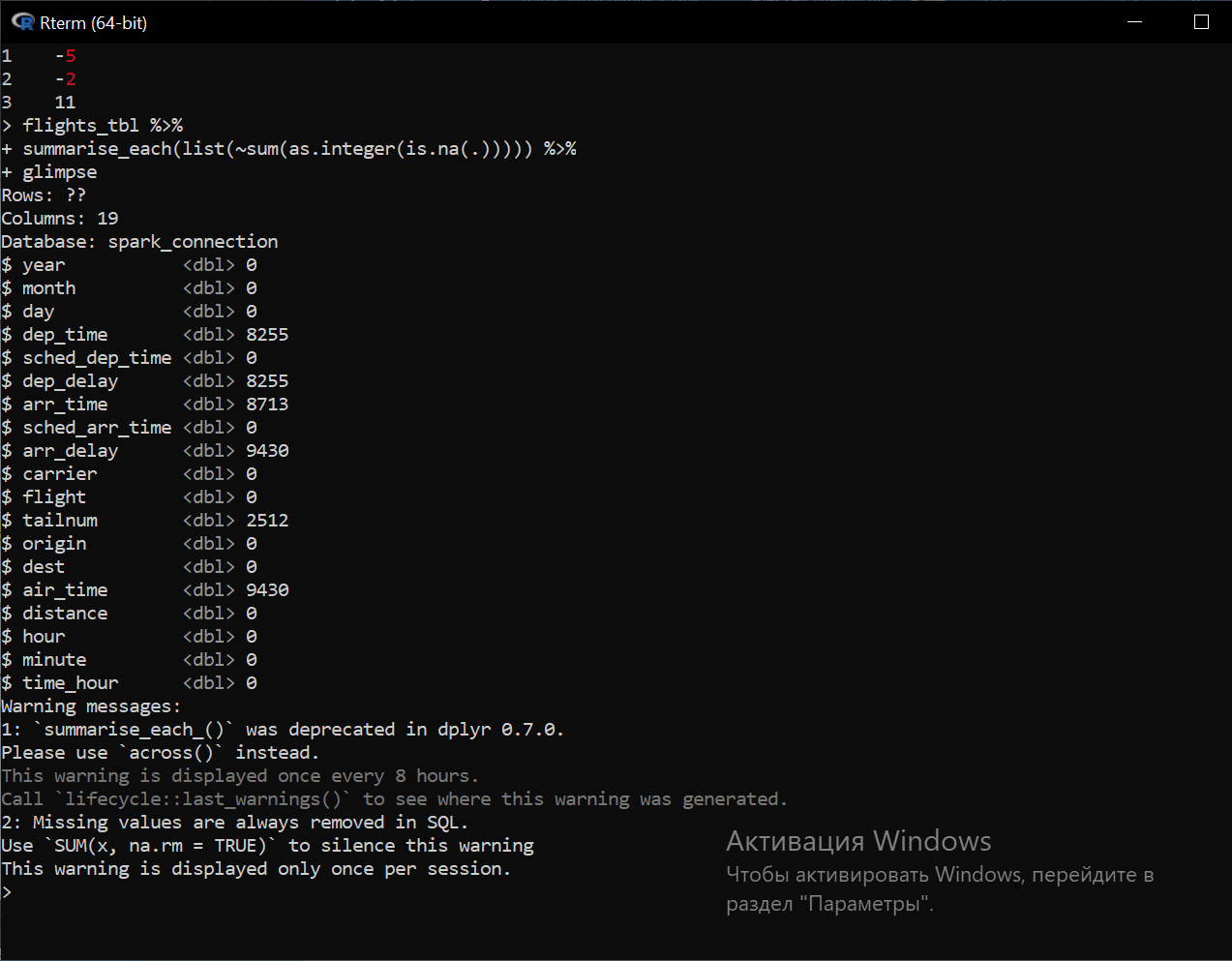


Щоб автоматично отримати ці значення зі списку служить Hive-

функція explode ():



Нижче підрахунок пропущених значень виконаний для всіх стовпців таблиці flights\_tbl за допомогою команди summarise\_each () з пакету dplyr в поєднанні з анонімною функцією, яка задає логіку обчислень:



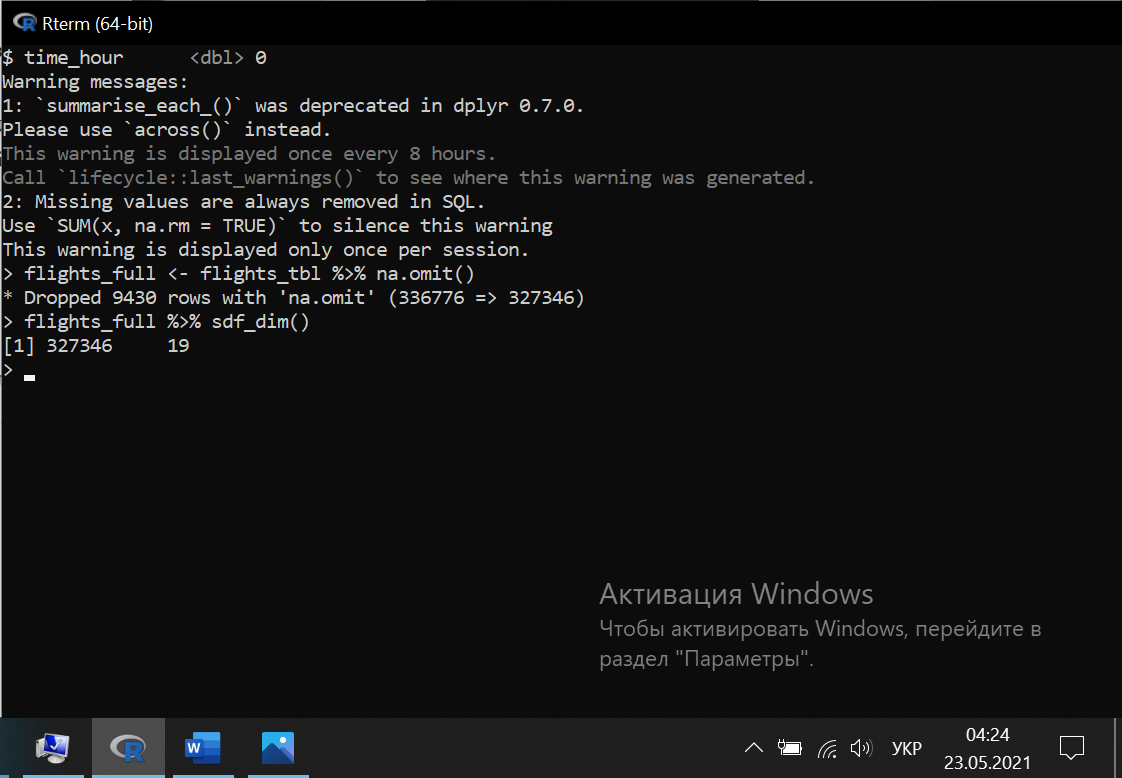
Яка максимальна кількість пропущених значень? Скільки це становить від

загального числа спостережень в таблиці?

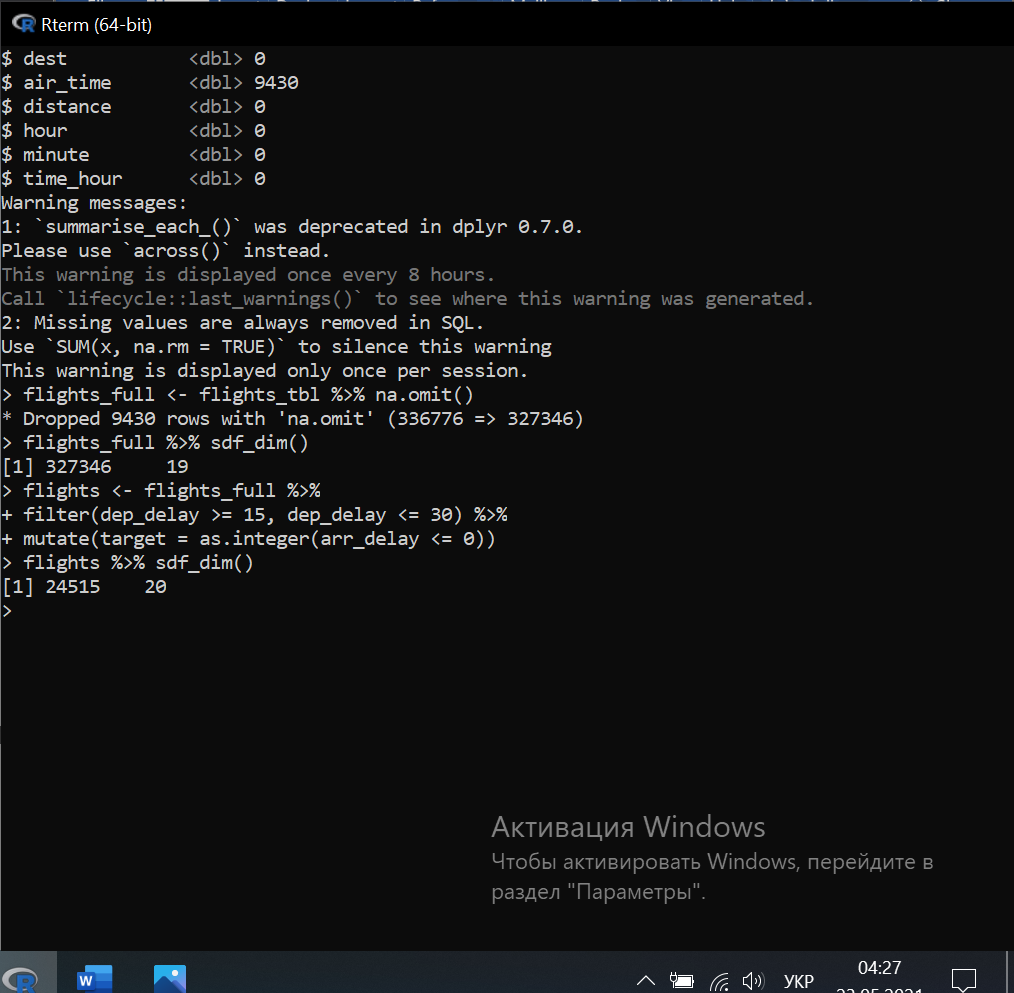
Максимальна кількість пропущених значень = 9430.

Це становить приблизно 2.8% (9430/336776)

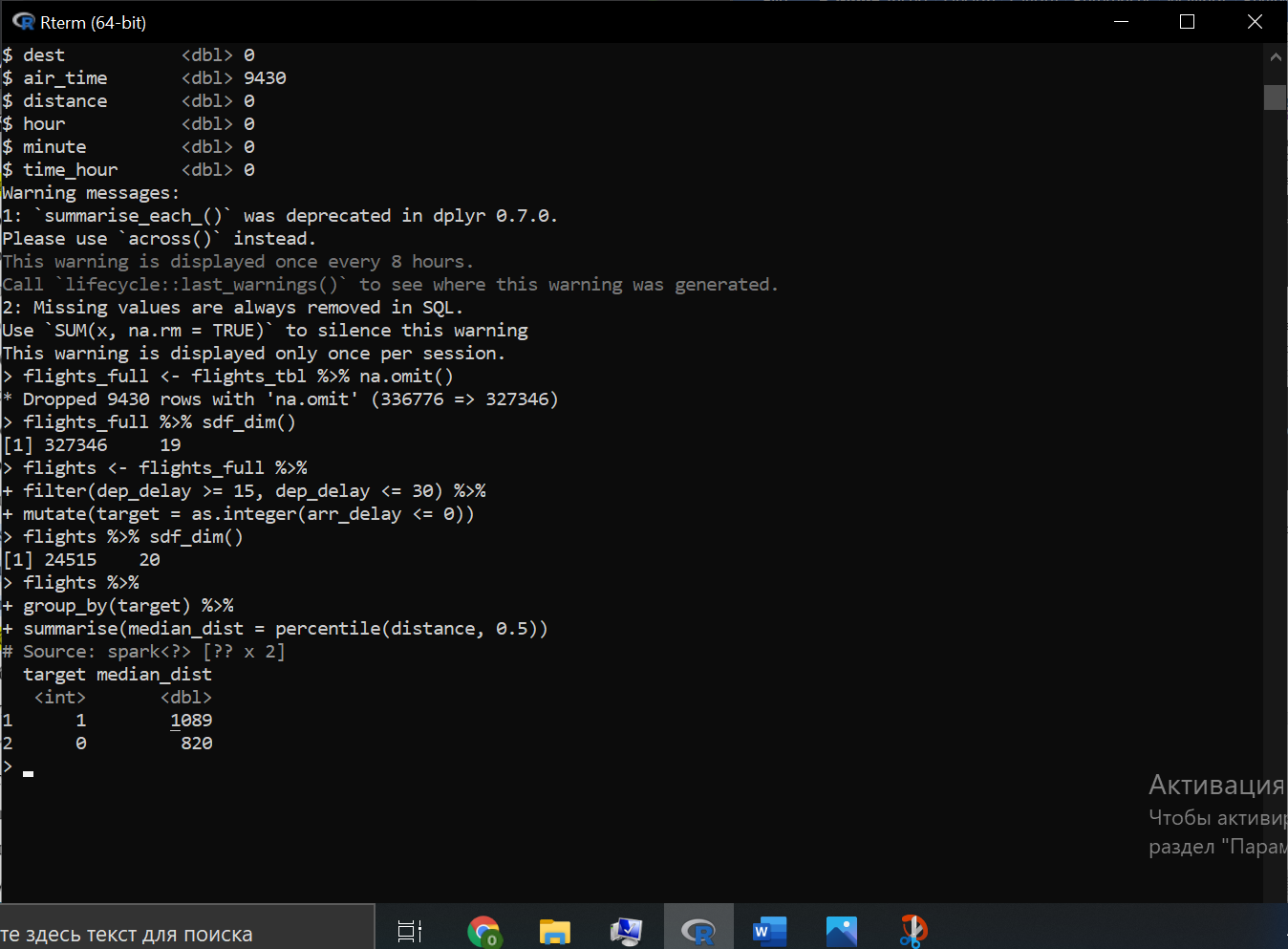
Якщо частка пропущених значень невелика, ми можемо видалити відповідні рядки з таблиці без особливого ризику вплинути на якість подальшого аналізу. Для цього можна скористатися базовою функцією na.omit ():



Оскільки нас цікавлять рейси, затримка яких склала від 15 до 30 хв (включно), далі нам потрібно відфільтрувати дані відповідним чином. Паралельно додамо новий стовпець target зі значеннями залежної змінної:



Розрахуємо медіанне значення цієї відстані для обох класів залежною змінною:



**Висновки**

В ході виконанні лабораторної роботи було освоєння використання Spark кластера на локальній машині для виконання обчислень для великих даних в середовищі R.