Были получены данные, которые содержат информацию об имени, фамилии, возрасте, электронной почте, зарплате и дате присоединения сотрудников.

В ходе предобработки данных были проделаны следующие шаги: поиск пропусков, дубликатов, проверка на соответствие данных ожидаемому типу, анализ аномальных значений.

1. В изначальных данных присутствуют пропуски: 1 в столбце «email», 22 в столбце «возраст», 12 в столбце «зарплата». Столбцы «имя», «фамилия», «дата присоединения» пропусков не содержат.
2. Все данные имеют тип object, что может говорить о некорректном заполнении исходной таблицы. Например, некоторые числовые значения были записаны как строковые, из-за чего произошло смешение типов. Для корректной работы на этом этапе столбец с датой присоединения был переведен в datetime, а столбец с возрастом и зарплатой в тип float.
3. На следующем шаге данные были проверены на дупликаты, которых не оказалось.
4. Столбцы с возрастом и зарплатой были проанализированы на адекватность значений. Возраст по новым требованиям должен быть больше или равен 18 и меньше 120, а зарплата не должна быть отрицательной.
5. После этих манипуляций были обработаны пропуски: строка без email была удалена, а пропуски в столбце с возрастом и зарплатой были заполнены средними значениями, чтобы не терять эти данные. Пропуски в столбце «дата присоединения» были также удалены.
6. В результате из 1000 строк сохранилось 848 – это около 85% от исходных данных.