|  |  |
| --- | --- |
|  | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

**Факультет «Информатика и системы управления»**

**Кафедра «Системы обработки информации и управления»**

Отчет по рубежному контролю №1

**«Технологии разведочного анализа и обработки данных»**

по дисциплине «Технологии машинного обучения»

Вариант №27

Выполнил:

студент группы ИУ5Ц-84Б   
Папин А.В.

подпись, дата

Проверил:

к.т.н., доц., Ю.Е. Гапанюк

подпись, дата

2024 г.

**СОДЕРЖАНИЕ ОТЧЕТА**

[1. Примечания: 3](#_Toc1)

[2. Дополнительные требования по группам: 3](#_Toc2)

[3. Листинг 4](#_Toc3)

[3.1. Подключение библиотеки и получение датасета 4](#_Toc4)

[3.2. Изучение данных 4](#_Toc5)

[3.3. Ящик с усами (Анализ выбросов) 5](#_Toc6)

[3.4. Пропущенные значения 5](#_Toc7)

[3.5. Дублирующие значения 6](#_Toc8)

[3.6. Удаление неинформативного признака 6](#_Toc9)

[3.7. Обработка категориальных признаков 6](#_Toc10)

[3.8. Создание нового признака 7](#_Toc11)

[3.8.1. Группировка по возрастам 7](#_Toc12)

[3.8.2. Группировка по уровня дохода 8](#_Toc13)

[3.9. Распределение городов и наличие болезней 10](#_Toc14)

[3.10. Доход в разрезе пола 10](#_Toc15)

[3.11. Распределения пола и наличия болезней 11](#_Toc16)

# **Примечания:**

Если в Вашем наборе данных отсутствуют данные, необходимые для решения задачи, создайте их искусственно. Например, если отсутствуют категориальные признаки, создайте категориальный признак на основе числового. Если отсутствуют пропуски, замените на пропуски часть значений в одном или нескольких признаках.

Также Вы можете дополнительно использовать датасеты, содержащие необходимые данные, например использовать дополнительный датасет, содержащий пропуски.

# **Дополнительные требования по группам:**

1. Для студентов группы ИУ5-64Б, ИУ5Ц-84Б - для произвольной колонки данных построить график "Скрипичная диаграмма (violin plot)".

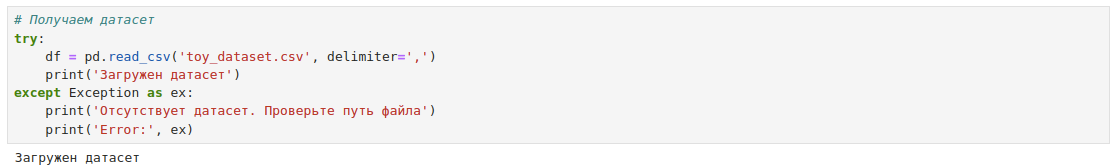
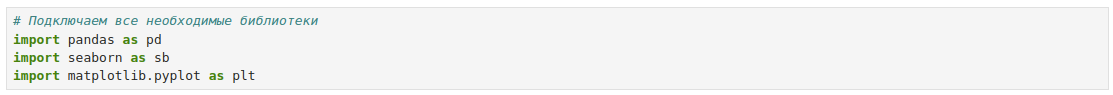
**Задача №4**

Для заданного набора данных постройте основные графики, входящие в этап разведочного анализа данных. В случае наличия пропусков в данных удалите строки или колонки, содержащие пропуски. Какие графики Вы построили и почему? Какие выводы о наборе данных Вы можете сделать на основании построенных графиков?

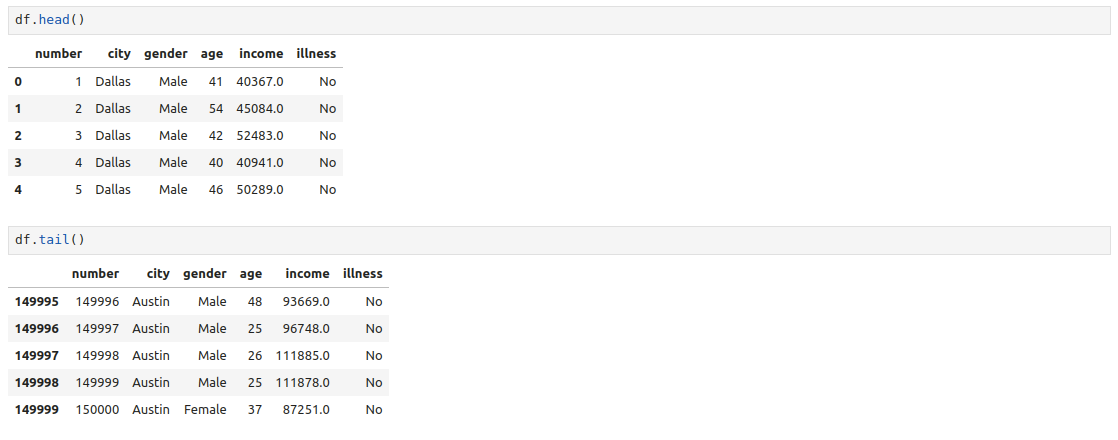
**Наборы данных:** https://www.kaggle.com/carlolepelaars/toy-dataset

# **Листинг**

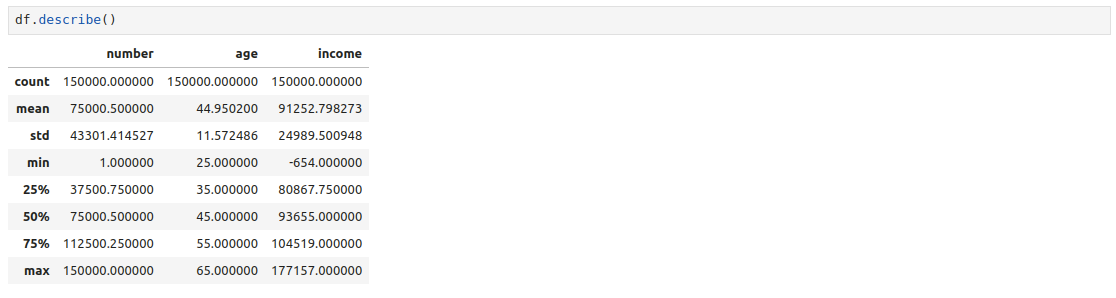
# **Подключение библиотеки и получение датасета**



# **Изучение данных**



Рассмотрим описательную статистику



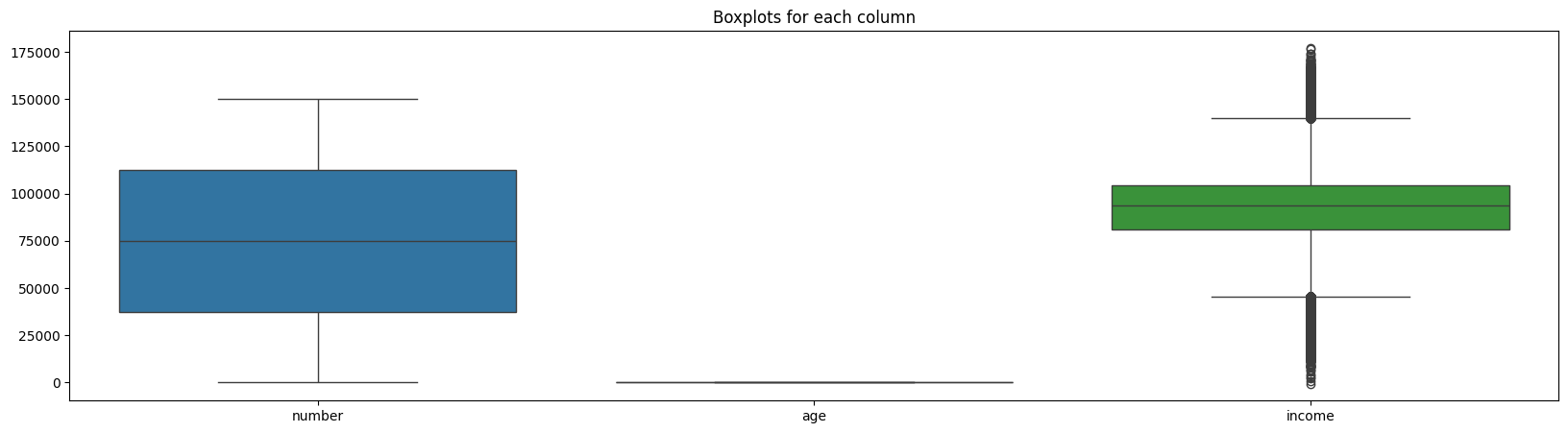
Анализ описательной статистики

number: Индексы распределены равномерно от 1 до 150000.

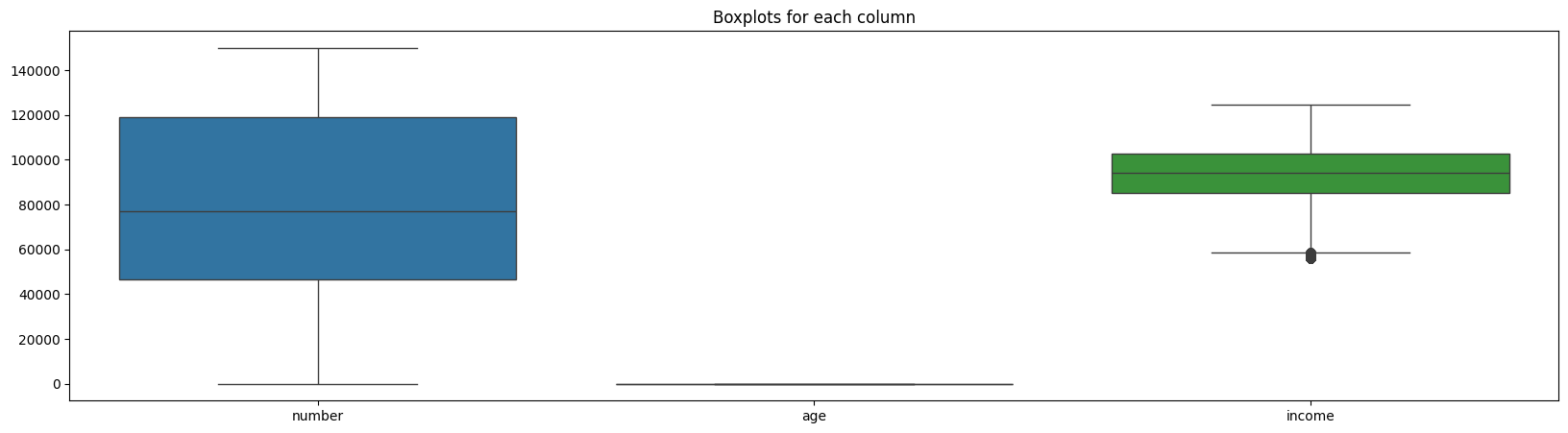
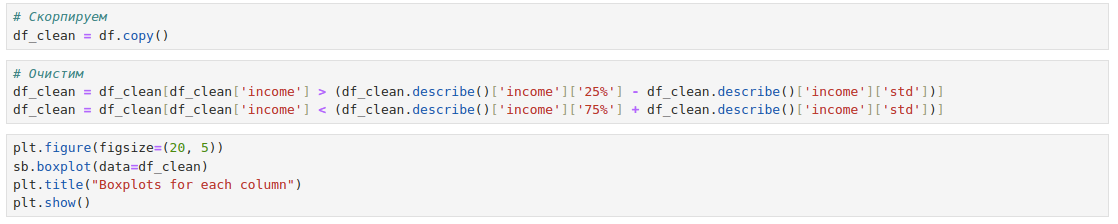
age: Средний возраст примерно 45 лет, с небольшим стандартным отклонением. Минимальный возраст 25 лет, максимальный - 65 лет.

income: Средний годовой доход около $91253, но с отрицательным минимальным значением, что может быть артефактом или ошибкой в данных.

# **Ящик с усами (Анализ выбросов)**

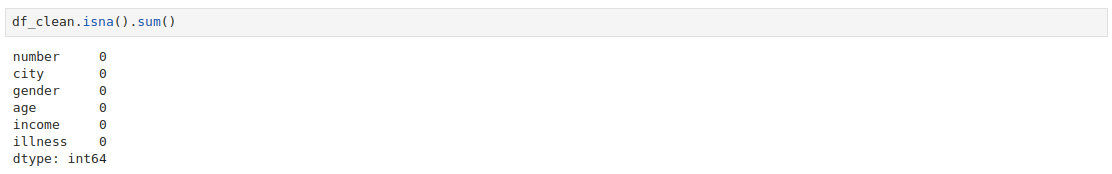


Виден огромный выброс у колонки income, устраняем.

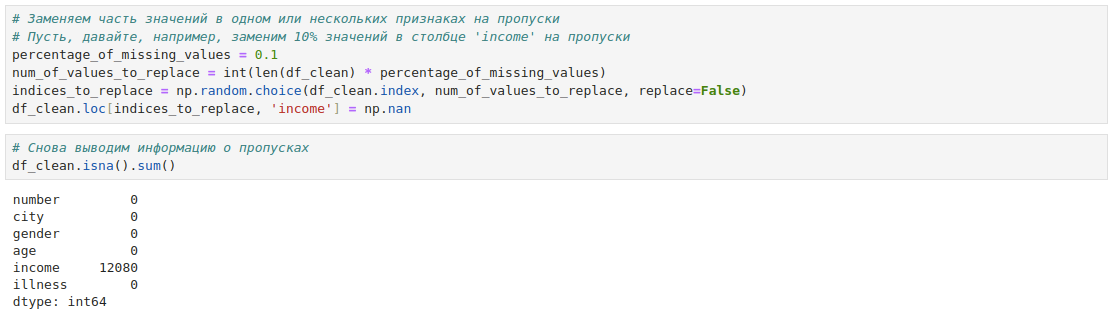


Получилось более менее нормально, с минимальным выбросом.

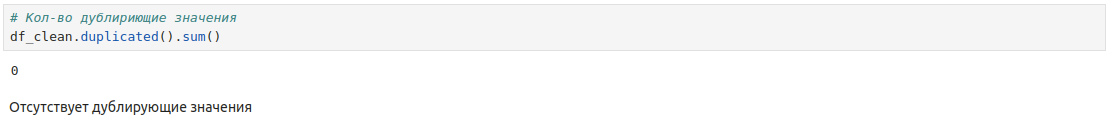
# **Пропущенные значения**



Отсутствует пропущенные значения. Давайте искусственно создадим их.

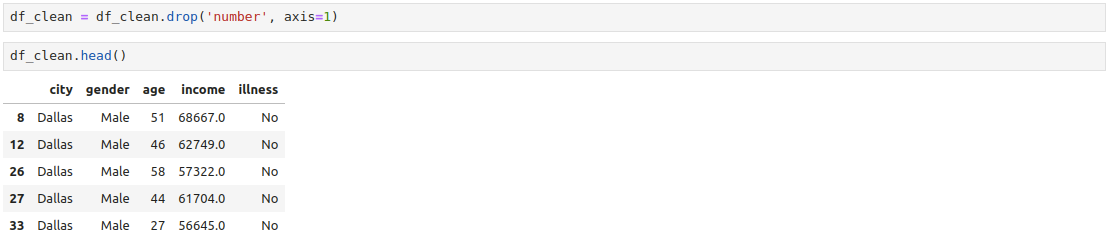


# **Дублирующие значения**



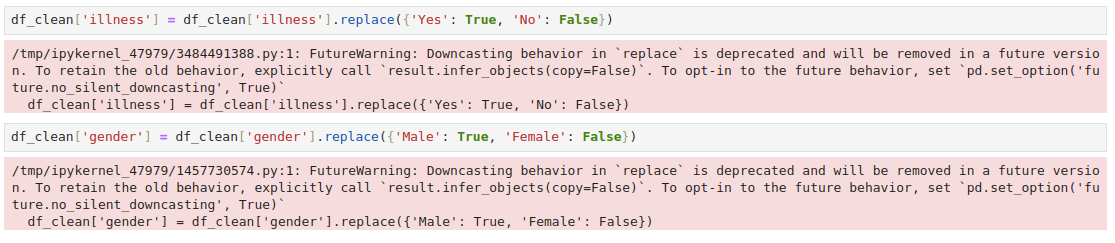
# **Удаление неинформативного признака**

Существует колонка - number, которая по сути является как индексация датафрейма, поэтому устраняем их.



# **Обработка категориальных признаков**

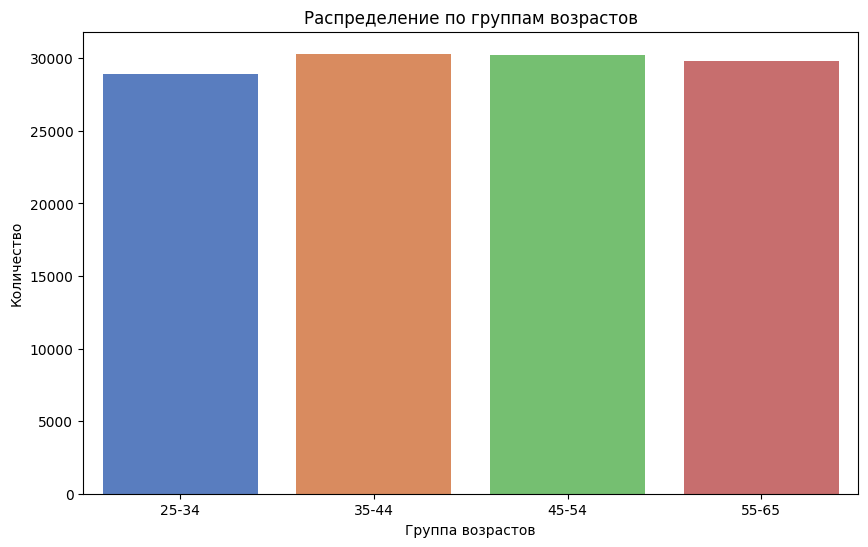
Можно закодировать названия города, но это потребуется для машинного обучения, поэтому колонку illness преобразуем в булевый тип.



# **Создание нового признака**

# **Группировка по возрастам**

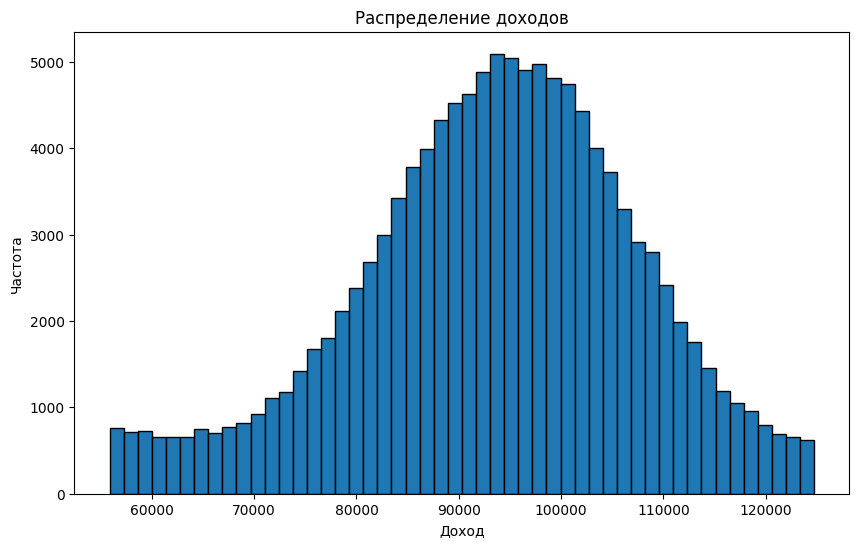
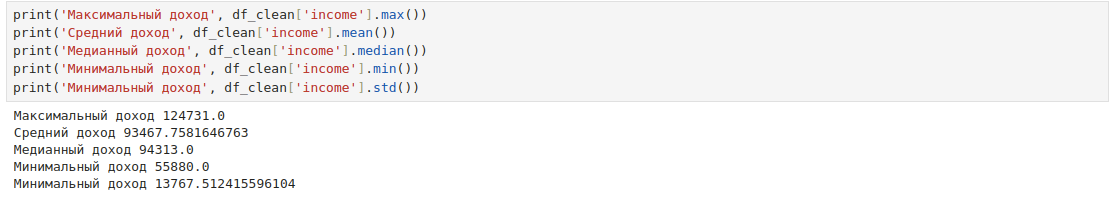
Создадим функцию, которая группирует данных по возрастам и создания новой колонки.



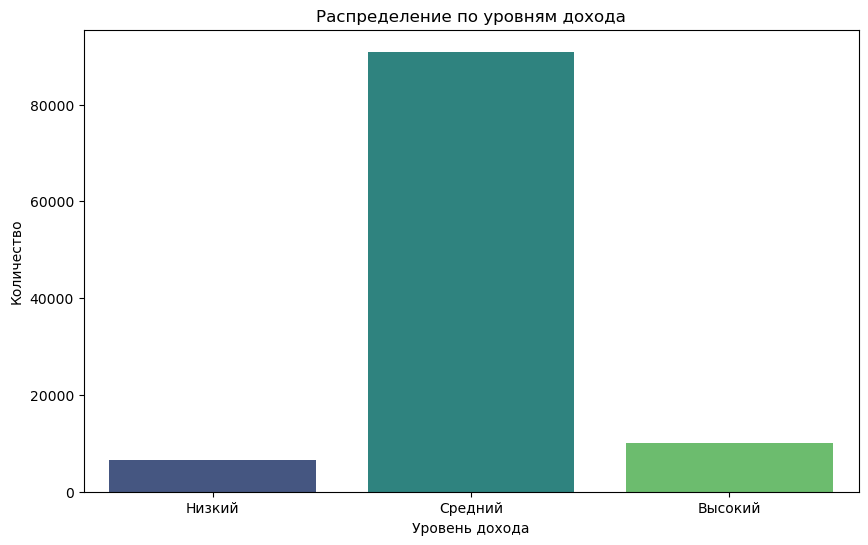
Как и видно, что все возрасты группы примерно равномерны одинаковы.

# **Группировка по уровня дохода**

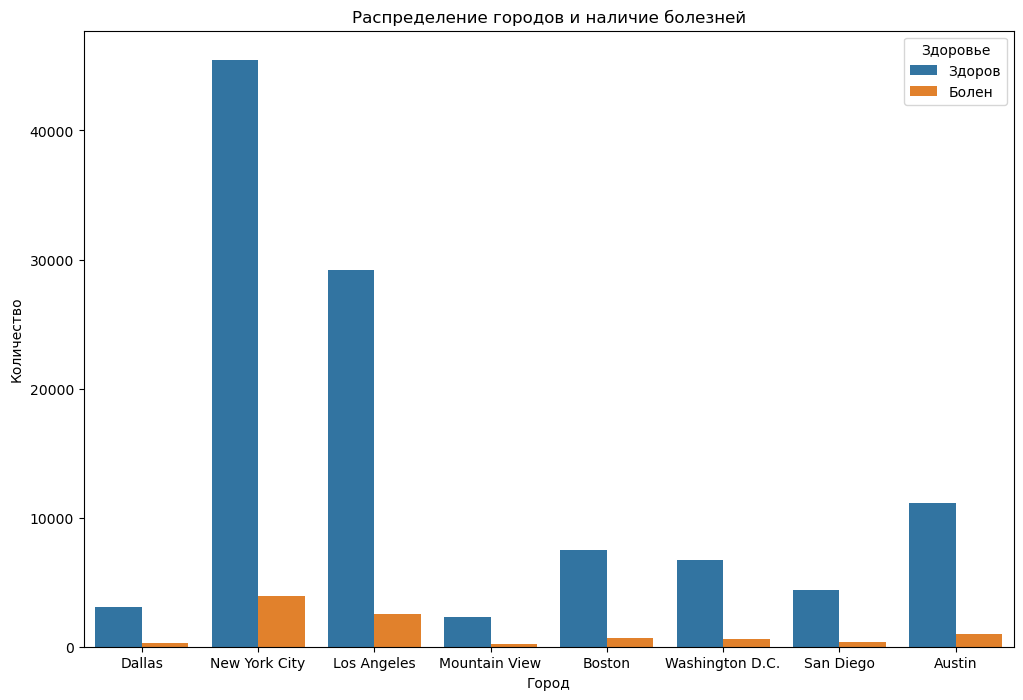
Создадим функцию, которая группирует данных по уровня дохода и по категории "Низкий", "Средний" и "Высокий" на основе заданных порогов.



Изобразим еще один график созданного нового признака.

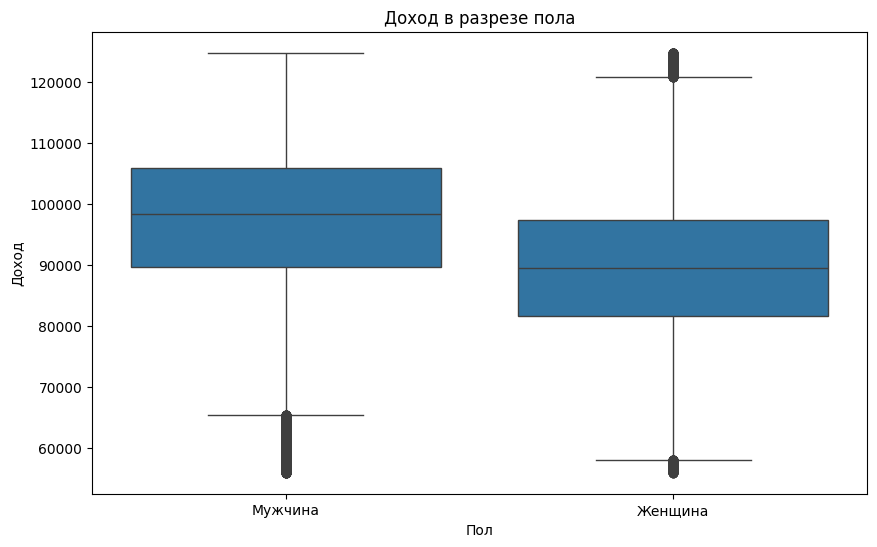
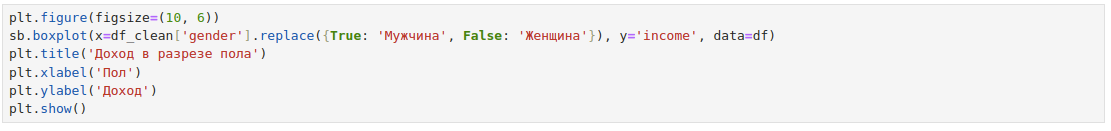


# **Распределение городов и наличие болезней**

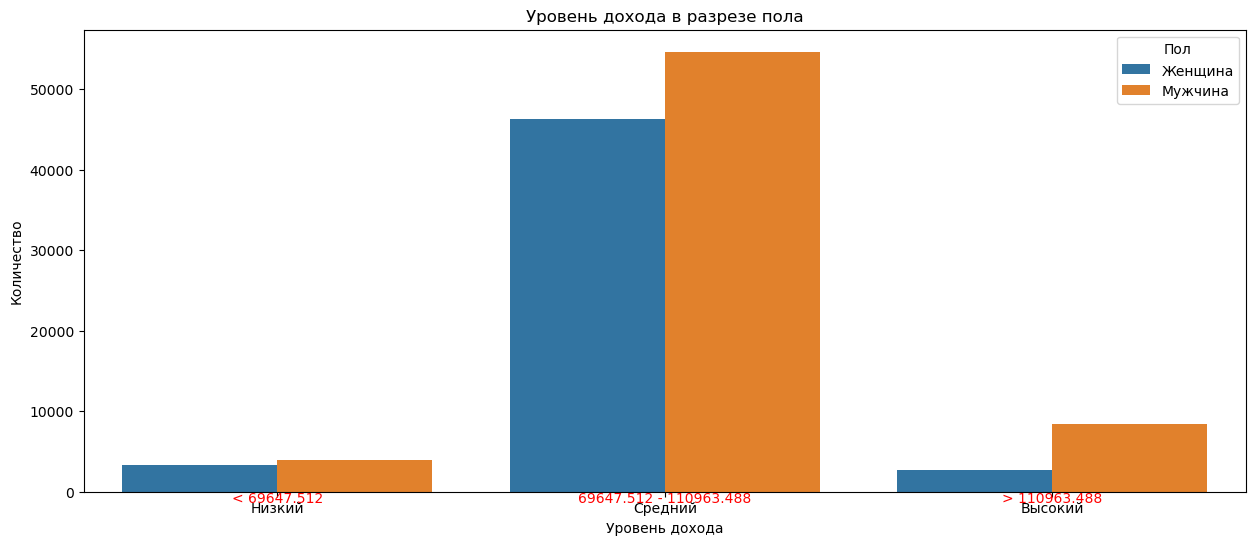


Как видим, что в густонаселенных городах мы часто сталкиваемся с риском по шансу схватиться с вирусными заболеваниями.

# **Доход в разрезе пола**

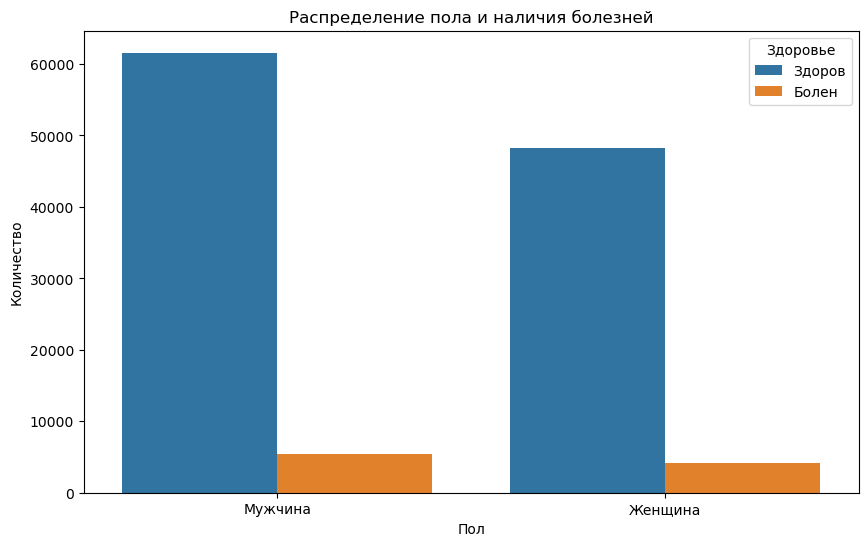
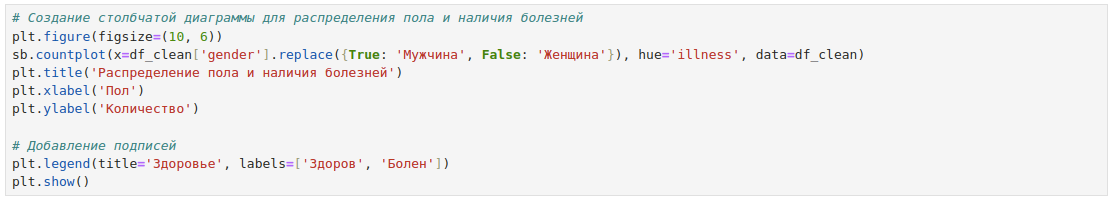


Будет проще рассмотреть по группировкам.

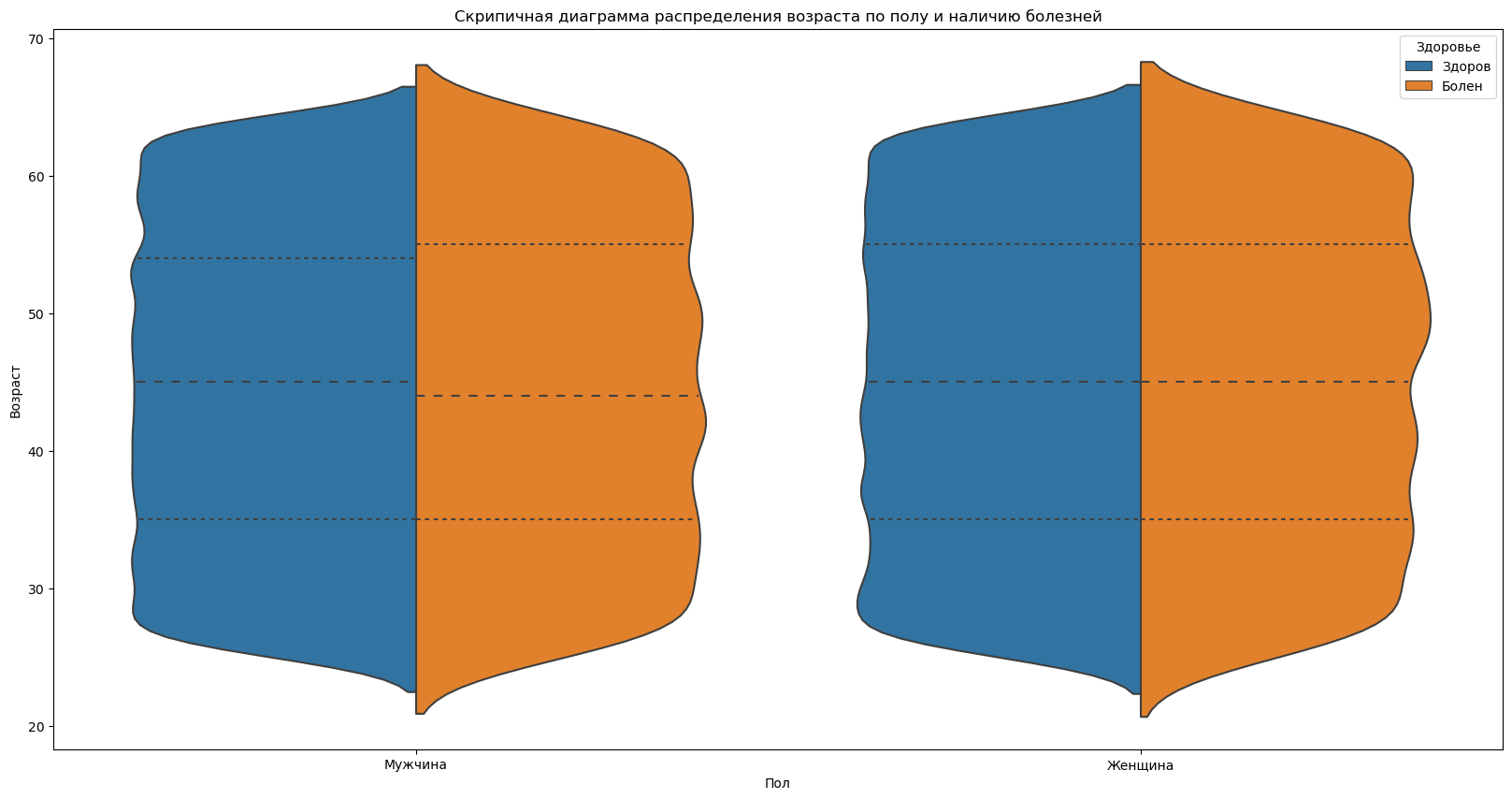


Сразу видно, что мужчины преобладает по большей заработной платы нежели представительницы прекрасного пола.

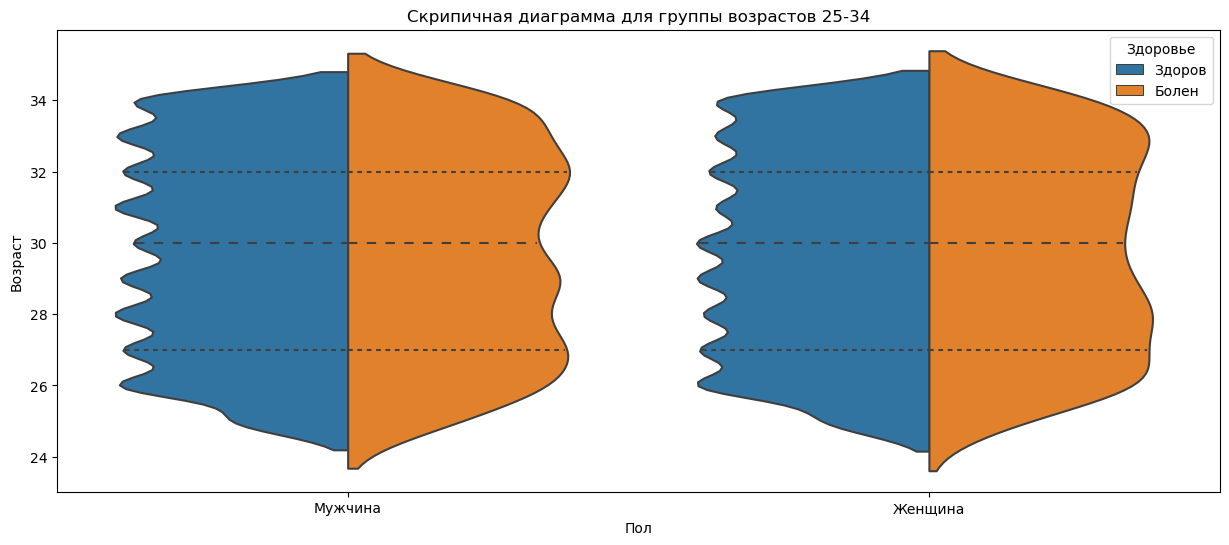
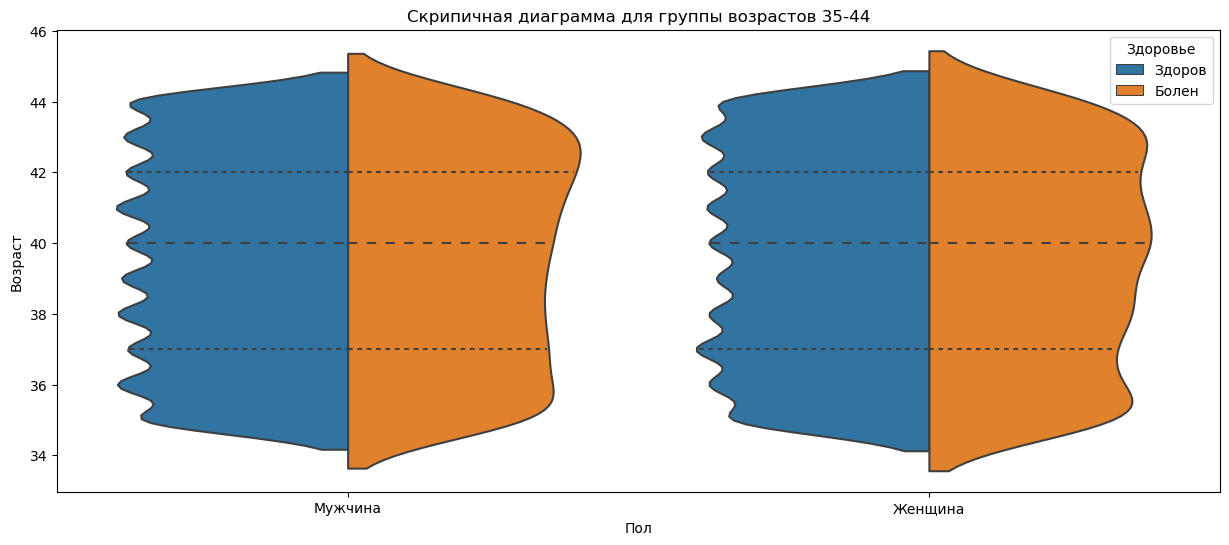
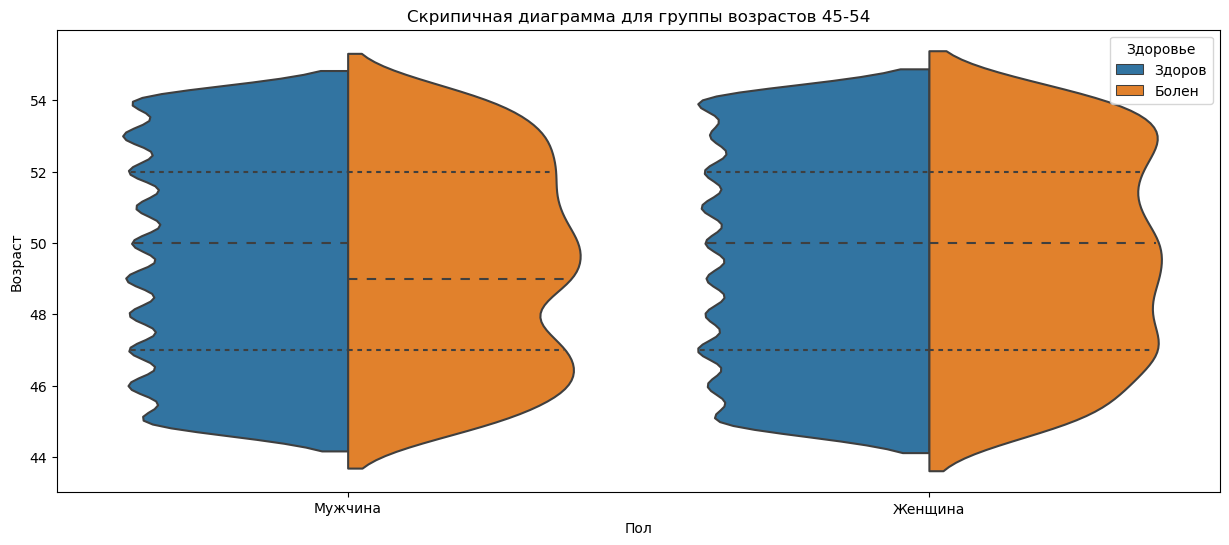
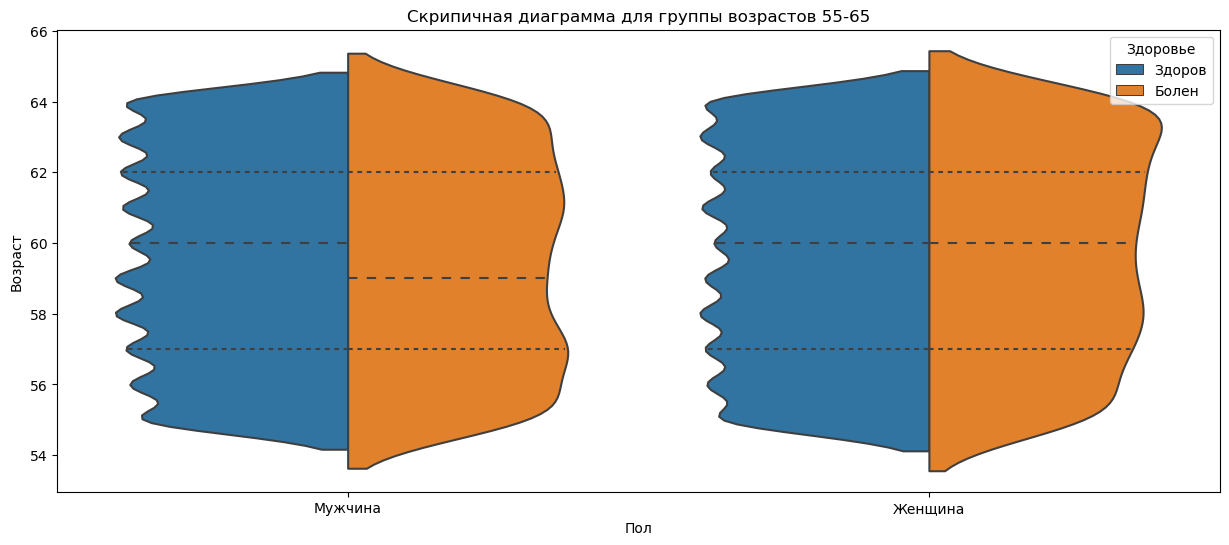
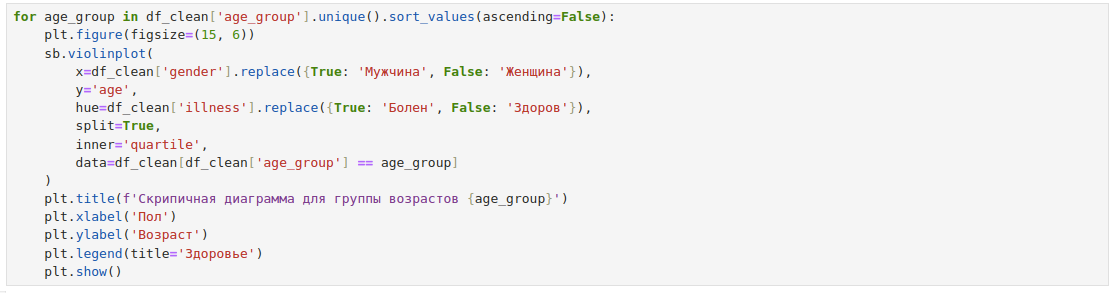
# **Распределения пола и наличия болезней**



Как и видим, что почти примерны одинаковы по уровня здоровья у мужчин и женщин. Лучше всего рассмотреть на скрипичной диаграммой.



Как и видим, что в молодом возрасте женщины реже заболевает нежели мужчин. А что если рассмотреть по группам возраста?



По графику можно примерно поставить оценку по уровня заболеваемости. В преклонных возрастах женщины заболевают чаще мужчин. В остальном примерно одинаковы получилось.