**Метрики анализа возраста в датасете**

1. **deviation\_from\_median:**

**Формула:** Отклонение от медианы = возраст - медианное значение возраста.

**Физический смысл:** Показывает, насколько возраст отклоняется от центрального значения в наборе данных. В нашем датасете медианное значение возраста составляет 67 лет.

**Пример:** Медианный возраст в датасете равен 67 лет. Если возраст определенного пациента 60 лет, то отклонение составит -7. Если возраст пациента 70 лет, то отклонение составит +3.

**Код:** data['deviation\_from\_median'] = data['age'] - data[‘age'].median()

**Применение показателя:** Используется для определения асимметрии распределения и для выявления значений, которые могут быть далеки от центра.

**2. age\_percentile:**

**Формула:** Процентный возраст = (количество людей с возрастом меньше возраста конкретного пациента / общее количество людей) \* 100.

**Физический смысл:** Для каждого уникального значения возраста в датасете этот показатель указывает, какой процент людей в датасете имеет возраст, который меньше этого конкретного значения.

**Пример:** Если у конкретного пациента этот показатель равен 70, это означает, что он старше 70% людей в датасете.

**Код:**

data['age\_percentile'] = data['age'].apply(lambda x: stats.percentileofscore(data['age'], x))

**Применение показателя:** Определяет ранговое место конкретного значения среди всех данных, что полезно в больших наборах данных, чтобы быстро понять относительное положение значения.

**3. age\_quantile:**

**Формула:** Разделение всех данных на четыре равные части.

**Физический смысл:** Позволяет классифицировать возраст в одну из четырех групп.

**Пример:** Если значение показателя для пациента равно 3, это означает, что он старше примерно 75% людей в выборке и одновременно младше самых старших 25%.

**Код:** data['age\_quantile'] = pd.qcut(data['age'], q=4, labels=[1, 2, 3, 4])

**Применение показателя:** Применяется для сегментации данных на равные группы что помогает в разделении данных на категории, основанные на квантилях и упрощает интерпретацию и анализ.

**4. is\_elderly:**

**Формула:** Если возраст больше порогового значения, то is\_elderly = 1; иначе is\_elderly = 0.

**Физический смысл:** Бинарная классификация, указывающая, превышает ли возраст определенный порог.

**Пример:** Если порог установлен на 70 лет и возраст человека 71 лет, то is\_elderly будет равно 1.

**Код:** threshold = 70

data['is\_elderly'] = (data['age'] > threshold).astype(int)

**Применение показателя:** Используется для быстрой бинарной классификации данных, особенно когда требуется отделить одну группу от другой на основе определенного порога.

**5. age\_density:**

**Формула:** Плотность вероятности возраста = значение оценки ядерной плотности в точке возраста (stats.gaussian\_kde).

**Физический смысл:** Оценка распространенности данного возраста в датасете.

**Пример:** Чем больше значение age\_density для определенного возраста, тем больше людей этого возраста в датасете. Например, если age\_density для 25 лет высок, это означает, что многие люди в выборке имеют такой возраст.

**Код:**

data['age\_density'] = data['age'].apply(lambda x: float(stats.gaussian\_kde(data[‘age'])(x)))

**Применение показателя:** Оценка плотности служит аналитическим инструментом для изучения распределения данных, выявления тенденций, пиковых значений и аномалий. Показатель используется в графической оценке, которая полезна при сравнении групп и заменяет гистограммы.

**6. age\_zscore:**

**Формула:** Z-оценка = (возраст - средний возраст) / стандартное отклонение возраста

**Физический смысл:** Относительное отклонение возраста от среднего значения, выраженное в единицах стандартного отклонения.

**Пример:** Если Z-оценка равна 2, это означает, что возраст пациента на два стандартных отклонения выше среднего возраста в датасете. Если Z-оценка отрицательная, возраст ниже среднего. Если она близка к нулю, возраст примерно равен среднему.

**Код:**

mean\_age = data['age'].mean() #средний возраст

std\_age = data['age'].std() # стандартное отклонение

data['age\_zscore'] = (data['age'] - mean\_age) / std\_age

**Применение показателя:** Z-оценка используется в статистике для определения отклонения от среднего, особенно когда нужно определить аномальные значения или сравнить разные наборы данных с разным масштабом и стандартными отклонениями.