Проблемы в данных

Пол	Возраст	Город	Доход	Образование	Вернул?
М	25	Мск	70000	A	1
Ж	31	Мск	100000000	C	1
М	10	Мск	40000	C	0
А	28	СПб	80000	B	0
М	23		40000	A	1
Ж	30	Екб	19973	B	1

Задача кредитного скоринга

Проблемы в данных

Пол	Возраст	Город	Доход	Образование	Вернул?
М	25	Мск	70000	A	1
Ж	31	Мск	100000000	C	1
М	10	Мск	40000	C	0
A	28	СПб	80000	B	0
М	23		40000	A	1
Ж	30	Екб	19973	B	1

Ошибка

Проблемы в данных

Пол	Возраст	Город	Доход	Образование	Вернул?
М	25	Мск	70000	A	1
Ж	31	Мск	100000000	C	1
М	10	Мск	40000	C	0
А	28	СПб	80000	B	0
М	23		40000	A	1
Ж	30	Екб	19973	B	1

Некорректное значение

Проблемы в данных

Пол	Возраст	Город	Доход	Образование	Вернул?
М	25	Мск	70000	A	1
Ж	31	Мск	100000000	C	1
М	10	Мск	40000	C	0
А	28	СПб	80000	B	0
М	23	?	40000	A	1
Ж	30	Екб	19973	B	1

Пропущенное значение

Проблемы в данных

Пол	Возраст	Город	Доход	Образование	Вернул?
М	25	Мск	70000	A	1
Ж	31	Мск	100000000	C	1
М	10	Мск	40000	C	0
А	28	СПб	80000	B	0
М	23		40000	A	1
Ж	30	Екб	19973	B	1

Выброс? Ошибка? Специальное значение?

Проблемы в данных

Пол	Возраст	Город	Доход	Образование	Вернул?
М	25	Мск	70000	А	1
Ж	31	Мск	100000000	С	1
М	10	Мск	40000	С	0
А	28	СПб	80000	В	0
М	23		40000	А	1
Ж	30	Екб	19973	В	1

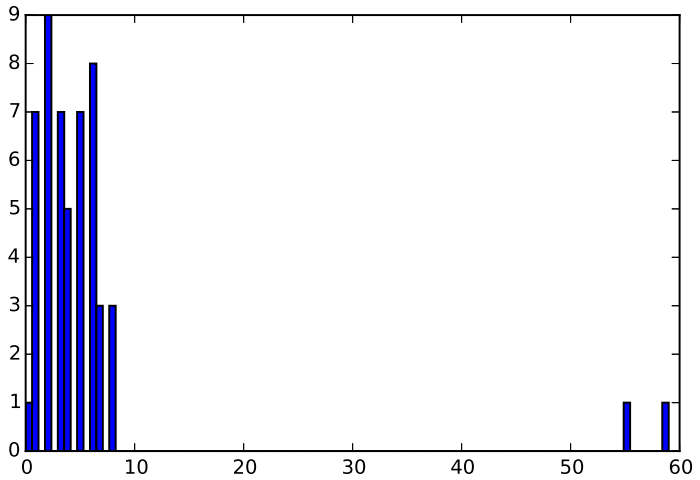
Подозрительная точность

Проблемы в данных

Пол	Возраст	Город	Доход	Образование	Вернул?
М	25	Мск	70000	A	1
Ж	31	Мск	100000000	C	1
М	10	Мск	40000	C	0
А	28	СПб	80000	B	0
М	23		40000	A	1
Ж	30	Екб	19973	B	1

Что означают эти буквы? Можно ли их сравнивать?

Фильтрация выбросов



Статистики:

- Первая квартиль Q_1: $\frac{1}{\ell}\sum_{i=1}^{\ell}[x_i <= Q_1] = 0.25$
- Третья квартиль Q_3: $\frac{1}{\ell}\sum_{i=1}^{\ell}[x_i <= Q_3] = 0.75$
- Интерквартильный размах: $\text{IQR} = Q_3 - Q_1$

Эвристика: выбросы лежат за пределами отрезка $[Q_1 - 1.5\ \text{IQR}, Q_3 + 1.5\ \text{IQR}]$.

Пропущенные значения

Пол	Возраст	Город	Доход	Образование	Вернул?
М	25	Мск	70000	A	1
Ж	31	Мск	?	C	1
М	10	?	40000	C	0
?	28	СПб	80000	B	0
М	23	?	40000	A	1
Ж	30	Екб	19973	B	1

Пропущенные значения

Пол	Возраст	Город	Доход	Образование	Вернул?
М	25	Мск	70000	A	1
Ж	31	Мск	49994.6	C	1
М	10	?	40000	C	0
?	28	СПб	80000	B	0
М	23	?	40000	A	1
Ж	30	Екб	19973	B	1

Пропуски в числовых признаках:

- **Замена на среднее**

Пропущенные значения

Пол	Возраст	Город	Доход	Образование	Вернул?
М	25	Мск	70000	A	1
Ж	31	Мск	40000	C	1
М	10	?	40000	C	0
?	28	СПб	80000	B	0
М	23	?	40000	A	1
Ж	30	Екб	19973	B	1

Пропуски в числовых признаках:

- Замена на среднее
- **Замена на медиану**

Пропущенные значения

Пол	Возраст	Город	Доход	Образование	Вернул?
М	25	Мск	70000	А	1
Ж	31	Мск	?	С	1
М	10	Мск	40000	С	0
М	28	СПб	80000	В	0
М	23	Мск	40000	А	1
Ж	30	Екб	19973	В	1

Пропуски в категориальных признаках:

- **Замена на самое популярное значение**

Пропущенные значения

Пол	Возраст	Город	Доход	Образование	Вернул?
М	25	Мск	70000	A	1
Ж	31	Мск	?	C	1
М	10	Пр	40000	C	0
Пр	28	СПб	80000	B	0
М	23	Пр	40000	A	1
Ж	30	Екб	19973	B	1

Пропуски в категориальных признаках:

- Замена на самое популярное значение
- **Замена на новое значение**

Пропущенные значения

Пол	Возраст	Город	Доход	Образование	Вернул?
М	25	Мск	70000	A	1
Ж	31	Мск	?	C	1
М	10	Екб	40000	C	0
Ж	28	СПб	80000	B	0
М	23	Мск	40000	A	1
Ж	30	Екб	19973	B	1

Пропуски в категориальных признаках:

- Замена на самое популярное значение
- Замена на новое значение
- **Случайный выбор из распределения значений**

Заключение

- В данных может быть много проблем: пропуски, выбросы, некорректные значения
- Выбросы можно обнаруживать с помощью интерквартильного размаха
- Пропуски можно исправлять путем замены на разумные значения