**Семинар 7-8.**

**Обязательно ответьте на вопросы теста на campus.fa.ru**

В файле «Кредиты\_2000\_0.tsv» представлены данные о выданных некоторым банком кредитах.

1. Загрузите данные в Excel, обращая внимание на форматы и разделители. Сколько столбцов у Вас получилось?
2. Проанализируйте типы переменных (непрерывные, дискретные, порядковые, номинальные). Для удобства работы имеет смысл сгруппировать стоблцы, содержащие числовые и нечисловые переменные, чтобы они шли по порядку (команда «Вырезать – Вставить вырезанные ячейки»).
3. Найдите в данных пропущенные значения и ошибки (неправильные форматы, лишние символы, языковые несоответсвия и т.п.). Используйте фильтры. **Исправьте ошибки.**

Работа с выбросами

1. Проведите анализ выбросов в этих данных. С помощью каких инструментов это можно сделать? Применим ли этот анализ к качественным признакам?

Убедитесь, что те признаки, которые Вы собираетесь анализировать, **действительно** являются числовыми!

Подсказка: чтобы проверить, является ли значение числом, попробйте применить к нему любую математическую операцию ( например, прибавить 1)

Подсказка 2: от некоторых ошибок можно избавиться с помощью функций Excel - ПЕЧСИМВ(), ЗНАЧЕН () и т. п.

Подсказка 3: чтобы посчитать количество выбросов используйте функцию СЧЕТЕСЛИ(), предварительно определив граничные значения (используйте функцию «СЦЕП()”)

1. Проанализируйте еще раз признак «Максимальный выданный кредит». Чем может быть обусловлен такой разброс данных? Как можно это исправить?

Работа с пропусками

Помните, что для симметричных распределений центр распределения одинаково хорошо характеризуется и средним, и медианой, а для несимметричных распределений (ассиметричность больше 1 по модулю) среднее значение неинформативно. Мода применяется для качественных признаков. Однако иногда надо заменять пропущенные значения не на основе статистических правил, а руководствуясь здравым смыслом. Предложите и обоснуйте, чем заполнять пропущенные значения в полях:

1. «Кредитный рейтинг».
2. «Годовой доход».
3. «Стаж на текущем месте работы».
4. «Срок с последнего нарушения условий кредитного договора».

**Заполните пропуски.**

Дублирующиеся строки

1. Идентифицируйте и удалите дублирующиеся строки в этом наборе данных (если они есть).

Синтетические признаки

Синтетические признаки используются для того, чтобы привести данные к единому масштабу, сделать распределение более симметричным или исходя из логики предметной области. Для приведения признаков к единому масштабу можно использовать z-преобразование или min-max преобразование. Синтетические признаки следует формировать после исправления ошибок, заполнения пустот и удаления дубликатов.

1. Если в задаче анализа данных о заемщиках известен размер ежемесячного платежа по кредиту и средний размер месячного дохода, то, как правило, полезно рассмотреть дополнительный признак - долю ежемесячного платежа по кредиту в месячном доходе. Постройте этот признак.
2. Постройте новые признаки, применив min-max преобразование к следующим полям:
   1. *Кредитный рейтинг*
   2. *Срок кредитной истории (лет)*
   3. *Срок с последнего нарушения кредитного договора (мес.)*
   4. *Количество кредитных карт*
   5. *Число нарушений кредитных договоров*
3. Постройте новые признаки, применив z- преобразование к следующим полям:
   1. *Размер кредита*
   2. *Годовой доход*
   3. *Ежемесячный платеж*
   4. *Текущий баланс кредитов*
   5. *Максимальный выданный кредит*

Анализ взаимосвязей

1. Постройте сводную диаграмму, характеризующую зависимость статуса кредита от стажа работы. Как влияет стаж работы на погашение кредита?
2. Постройте диаграмму рассеяния для признаков «Годовой доход» и «Кредитный рейтинг». Есть ли здесь взаимосвязь?
3. Сравните средний кредитный рейтинг (до устранения пропусков) клиентов, погасивших кредиты и нет. Можно ли сказать, что кредитный рейтинг хороших заемщиков (статус погашен) в среднем выше?