**Трубчик А.И.**

План подготовки и очистки данных - применительно к R, основано на опыте подготовки данных German Credit Data c искусственно добавленными 10% пропусков.

**Загрузка данных**

Загрузить данные в data frame с помощью функций read.table/read.csv. Категориальные переменные представить в виде factor. Привести данные типам R, т. е. числовые должны быть в формате integer/number.

**Проверка на ошибки и выбросы**

Количественные данные должны иметь адекватные значения, например, возраст человека не может быть отрицательным. Визуальную проверку на выбросы можно осуществить с помощью диаграммы размаха (функция boxplot). Чтобы уменьшить влияние выброса, его значение можно ограничить, например, для возраста установить лимит в 60 лет.

**Преобразование данных**

Некоторые категориальные данные могут содержать параметры, которые можно выделить в отдельную категорию. Например, данные могут содержать переменную, включающую информацию о семейном положении и пол человека, которые разумно разделить на две — отдельно пол, отдельно семейное положение.

Если данные отличаются размерностью, то следует привести их к одному виду - библиотека deducorrect.

**Восстановление данных**

Исходные данные могут иметь пропущенные значения. Для восстановления количественных данных можно использовать линейную регрессию, среднее значение (по общей выборке или по схожим данным) и другие способы в зависимости от вида этих данных. Для категориальных переменных можно применить логистическую регрессию.

Автоматизировать процесс восстановления данных возможно с помощью библиотек mi, mice, VIM и других.

**Подготовка для классификации**

Провести корреляционный, дисперсионный, регрессионый и/или факторный анализы, чтобы определить взаимосвязь между переменными и сократить их число.

Например, с помощью линейной регрессии и функции vif можно проверить переменные на мультиколлинеарность.

Можно проверить корреляцию между количественными переменными: corrgram из библиотеки corrgram, функцией cor, графиком scatterplot. Если они слишком коррелированы, то можно попытаться вычленить их составляющие. Однако наличие корреляции не означает, что одна из переменных является причиной влиянию на другую — обе могут быть подвержены влиянию третьего фактора.

Например, сумма кредита и срок кредитования имеют высокое значение коэффициента корреляции. Разделив сумму на срок, получим ежемесячный платеж (без учета процента банка). Корреляция между ним и сроком кредитования будет минимальной.

Для категориальных данных следует использовать другие методы ([ссылка](http://stats.stackexchange.com/questions/119835/correlation-between-a-nominal-iv-and-a-continuous-dv-variable/124618" \l "124618)), например, библиотеку hetcor.