Семинар 9

На этом семинаре мы будем работать с известной базой данных (http://campus.lakeforest.edu/frank/FILES/MLFfiles/Bio150/Titanic/TitanicMETA.pdf) по пассажирам "Титаника" (она часто используется в курсах по эконометрике и машинному обучению, но представляет не только статистической, но и содержательный интерес).

Переменные:

PassengerId - id пассажира

Survived - бинарный показатель, выжил пассажир или нет (1 - выжил, 0 - не выжил)

Pclass - класс пассажира

Name - имя пассажира

Sex - пол пассажира

Age - возраст пассажира

SibSp - число родных братьев/сестер пассажира на борту корабля (или супругов)

Parch - число родителей пассажира на борту корабля

Ticket - номер билета

Fare - стоимость билета

Cabin - каюта

Embarked - порт, в котором пассажир взошел на палубу корабля

Загрузка и предварительная обработка

1. Загрузите базу данных из файла `Titanic.csv` .
2. Загрузите базу данных так из файла еще раз, но так, чтобы столбец PassengerId был идентификатором, то есть номером строки (*index)*.
3. Удалите из базы строки с пропущенными значениями и сохраните изменения в самой базе.

Описание базы данных

3. Выведите сводную информацию по базе данных: какие переменные в ней есть, какого они типа + сколько заполненных наблюдений в каждой столбце.
4. Выведите сводную статистическую информацию по каждому количественному показателю в базе (описательные статистике).
5. Постройте гистограмму для переменной *Возраст* (*Age*), сделайте ее красного цвета, подпишите оси и добавьте заголовок графика.
6. Выведите описательные статистики для столбца *Стоимость билета* (*Fare*).

Работа со строками и столбцами базы

7. Выведите названия столбцов в базе данных в виде списка (объект типа *list*).
8. Переименуйте столбец с классом пассажира из **Pclass** в **Class**.
9. Выберите из базы данных все строки, которые соответствуют пассажирам женского пола, и сохраните их в новую базу `female`.
10. Выберите из базы данных все строки, которые соответствуют выжившим пассажирам мужского пола младше 32 лет, и сохраните их в базу `Ymale`.
11. Выберите из базы данных все строки, которые соответствуют пассажирам 1 или 2 класса.
12. Выберите из базы данных все строки, которые соответствуют выжившим пассажирам 1 или 2 класса.
13. Добавьте в датафрейм столбец *Female*, состоящий из значений 0 и 1, где 1 соответствует пассажирам женского пола.

Группировка

1. Выведите на экран все уникальные значения в столбце *Embarked*.
2. Сгруппируйте строки в датафрейме в соответствии со значениями переменной *Survived* и выведите средние значения всех количественных переменных по группам.
3. Сгруппируйте строки в датафрейме в соответствии со значениями переменной *Sex* и сохраните в отдельный датафрейм таблицу со средними и медианными значениями переменной *Age* по группам (мужчины и женщины).

Выгрузка базы в файл

1. Приведите все названия столбцов в датафрейме к нижнему регистру и сохраните изменения.
2. Выгрузите итоговый датафрейм в файл `Titanic-new.csv`.