Лабораторная работа №5
Практическая статистика и визуализация с Python

Пояснение к заданию:

1. Задание требуется выполнять jupyter notebook.
2. Для подключения к базе данных clickhouse можно использовать библиотеку clickhouse_driver.
3. Для подключения к базе можно использовать следующий код

```
from clickhouse_driver import Client

client = Client('oleg.orbita.work', port = 9000, user=student_ml,
password=CGRV8zemLGgsfdz7Uc6ZJeisGTcPQa, settings={'use_numpy':
True})
```

Задание

1. Загрузите данные из таблицы house_train, которая находится в базе students_datas.
2. Приведите описание датасета:
   2.1. Сколько данных в датасете?
   2.2. Сколько параметров? Выведите список всех параметров.
   2.3. Есть ли категориальные признаки? Перечислите / выведите их.
   2.4. Выведите первые пять строчек DataFrame.
3. Проверьте, есть ли пропуски и повторы в данных.
   3.1. Удалите повторы
   3.2. Удалите столбцы в которых пропущено более 15% данных
4. Постройте гистограмму параметра SalePrice, Подчиняется ли распределение нормальному?
5. Построить коробочную диаграмму (ящик с усами) признака SalePrice всех домов в данных.
6. Постройте Гистограммы и Боксплоты по группам:
   6.1. кондиционером ('CentralAir') и без кондиционера
   6.2. цены продажи домов (параметр 'SalePrice'), сгруппированные по размеру гаража (параметр 'GarageCars')
7. Постройте гистограмму частот:
   7.1. частот размеров гаража
   7.2. частот центрального кондиционирования
8. Рассчитайте долю домов
   8.1. с продажной ценой между 25-м процентилем и 75-м процентилем .
   8.2. Рассчитайте долю домов с общей площадью в квадратных футах от 25-го процентиля до 75-го процентиля .
9. Получите ковариационную матрицу для всех данных DataFrame и используйте анализ тепловой карты, Выведетите 10 параметров с наибольшей корреляцией с SalePrice,