```
import os
# c:/aaa/_R/_R3/Ames_dataset/
os.chdir("c:/aaa/_R/_R3/Ames_dataset/")

import pandas as pd
import numpy as np

# https://pandas.pydata.org/pandas-docs/stable/generated/pandas.read_csv.html
# Полезные параметры:

#     sep -- задаёт символ-разделитель полей в файле (по умолчанию разделитель запятая);
#     names -- список названий колонок, если он не задан в файле;
#     index_col -- номер колонки с индексом.
#     decimal -- символ-разделитель для знаков после запятой.
#         decimal=b',',

AH = pd.read_csv('AmesHousing.txt', sep="\t", header = 0, index_col=False)



AH.head()


print(AH.shape)
print(len(AH))


AH.dtypes


AH.describe(include='all')


import matplotlib
import matplotlib.pyplot as plt

# выбор темы картинок на Ваш вкус. Необязательная строчка.
matplotlib.style.use('ggplot')
# Эта строчка нужна для того, чтобы картинки отображались в ячейках
%matplotlib inline


#  https://matplotlib.org/api/_as_gen/matplotlib.pyplot.hist.html


AH['SalePrice'].hist();


AH['SalePrice'].hist(bins=60);


AH['SalePrice'].hist(bins=60, normed=1);
```

```
print(plt.style.available)


matplotlib.style.use('seaborn-deep')

AH['SalePrice'].hist(bins=60, normed=1);



np.log(AH['SalePrice']).hist(bins=45, normed=1);



from scipy.stats.kde import gaussian_kde

from numpy import linspace,hstack
from pylab import plot,show,hist


#      недостаток!!
#  отсутствует sheather jones bandwidth
#  Scott Silverman normal density reference


# создадим функцию
# создадим функцию
my_density = gaussian_kde(AH['SalePrice'])
#  my_density = gaussian_kde(AH['SalePrice'], bw_method = 5)
#  my_density = gaussian_kde(AH['SalePrice'], bw_method = 1)
#  my_density = gaussian_kde(AH['SalePrice'], bw_method = 0.1)


# график
x = linspace(min(AH['SalePrice']), max(AH['SalePrice']),1000)
plot(x, my_density(x),'g') # distribution function

hist(AH['SalePrice'], normed=1, alpha=.3) # histogram
show()


plot(x, my_density(x),'r') # distribution function



# По идее, вызов должен быть такой: df.groupby('Status')['Length'].hist(alpha=0.6)
# Но из-за бага https://github.com/pandas-dev/pandas/issues/10756
# приходится делать дополнительный вызов plot
AH.groupby('MS Zoning')['SalePrice'].plot.hist(alpha=0.6)
# Добавляем легенду
plt.legend();
```

```
AH.groupby('MS Zoning')['SalePrice'].plot.hist(normed=1, alpha=0.6)
# Добавляем легенду
plt.legend();



ax = AH.boxplot(column='SalePrice', by='MS Zoning')
# Хак для того, чтобы исправить наезжающие заголовки графика
ax.get_figure().suptitle('')


print (AH['MS Zoning'].value_counts())



#  разброс данных
#  дисперсия
#  стандартное отклонение
#  разброс
#  IQR


# города россии


# import os
# c:\aaa\temp_py\Shad_Python_01_2\
os.chdir("c:/aaa/temp_py/Shad_Python_01_2/")

# import pandas as pd
# import numpy as np



town = pd.read_csv('town_1959_2.csv', encoding='cp1251', index_col=u'номер')
town.head()



print(town)


town.describe()


len(town[town['население'] < 52.925199])/len(town)*100
```

```
#  town['население'].median()

town_2 = town.iloc[2:1004,:]
#  print(x_2)

town_2.describe()

len(town_2[town_2['население'] < 44.997904])/len(town_2)*100


town[u'население'].hist()


x = np.log10(df[u'население'])
pd.Series(x).hist()


pd.Series(x).hist(bins=45)
```