|  |
| --- |
|  |
| МИНОБРНАУКИ РОССИИ |
| Федеральное государственное бюджетное образовательное учреждение  высшего образования  **«МИРЭА - Московский технологический университет»**  **РТУ МИРЭА** |

Институт Информационных Технологий

Кафедра Прикладной Математики (ПМ)

**ПРАКТИЧЕСКАЯ РАБОТА № 4**

|  |  |
| --- | --- |
| Выполнил студент группы ИКБО-08-19  Борисов А.В. | (\_\_\_\_\_\_\_\_\_\_)  *подпись* |
| Принял Ассистент кафедры ПМ  Высоцкая А.А. | (\_\_\_\_\_\_\_\_\_\_)  *подпись* |
| Практическая работа выполнена | « » 2022 г. |
| «Зачтено» | « » 2022 г. |

Москва 2022

1. Загрузить данные из файла “insurance.csv”

import pandas as pd

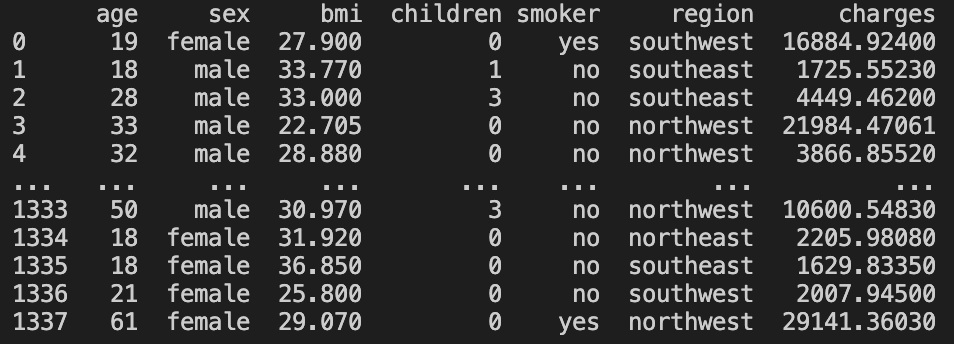
#'age,sex,bmi,children,smoker,region,charges'

def getData():

data = pd.read\_csv("insurance.csv", sep=',')

return data

print(getData())

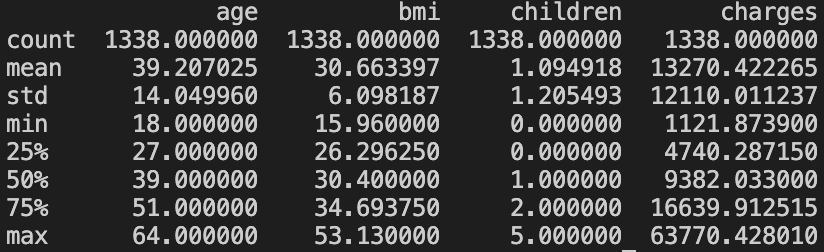
*.*

1. С помощью метода describe() посмотреть статистику по данным. Сделать выводы.

from firstTask import getData

inshuranceData = getData()

print(inshuranceData.describe())



count – всего значений

mean – среднее значение

std – стандартное отклонение

min – минимальное значение

25% - Q1 (первый квартиль)

50% - Q2 (второй квартиль)

75% - Q3 (третий квартиль)

max – максимальное значение

1. Построить гистограммы для числовых показателей. Сделать выводы.

import seaborn as sns

import matplotlib.pyplot as plt

import firstTask

inshuranceData = firstTask.getData()

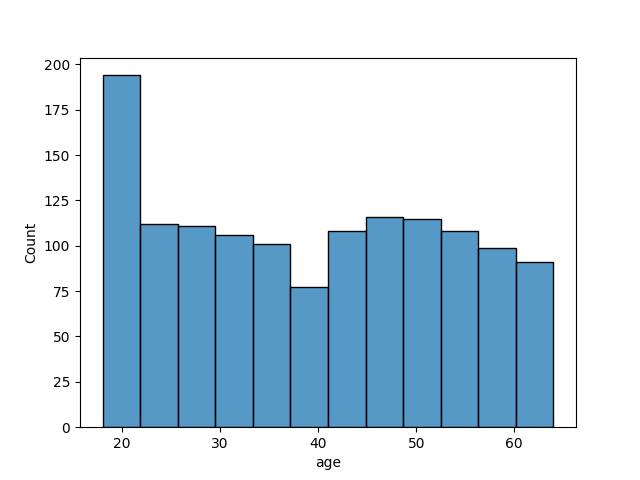
#sns.histplot(data=inshuranceData['age'])

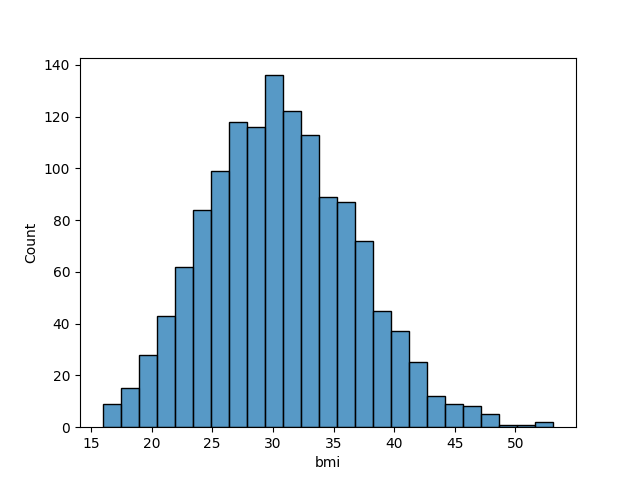
#sns.histplot(data=inshuranceData['bmi'])

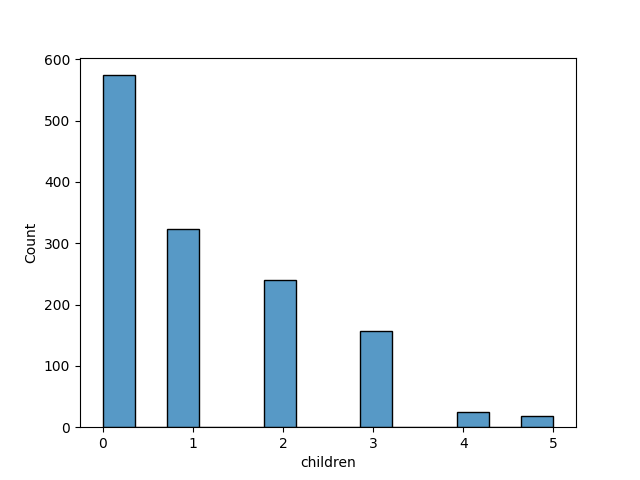
#sns.histplot(data=inshuranceData['children'])

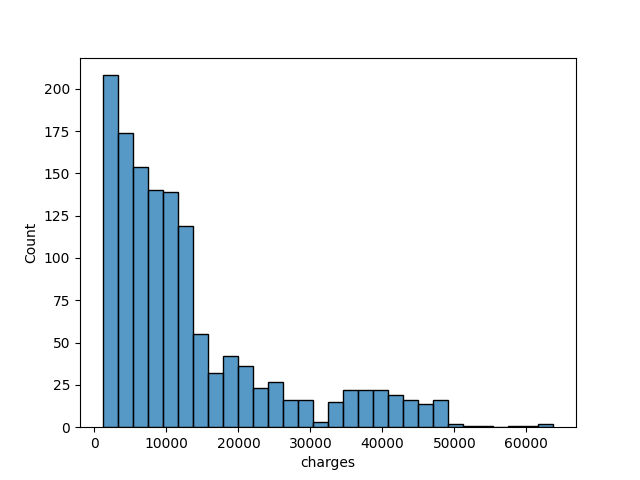
sns.histplot(data=inshuranceData['charges'])

plt.show()









Гистограммы отображают количественные соотношения величин age, bmi, children и charges.

**Ages**. Из данной диаграммы можно сделать вывод, что в данной выборке самая большая группа – это люди ± 20 летнего возраста, однако основная масса все-равно старше.

**Bmi**. Данная гистограмма может быть похожа на график плотности нормального распределения. Основная масса сосредоточена на показателях от ±26 единиц (Q1), что в принципе гласит, что большая часть людей из данной выборки уже находится практически на начальных стадиях ожирения…

**Children**. Тут все проще, большая часть – не имеют детей, и количество людей, которые имеют более одного ребенка уменьшается с каждым последующим ребенком.

**Charges**. Данная гистограмма гласит, что большинство имеют сумму выплат менее 15000, и лишь единицы более 50000.

1. Найти меры центральной̆ тенденции и меры разброса для индекса массы тела (bmi) и расходов (charges). Отобразить результаты в виде текста и на гистограммах (3 вертикальные линии). Добавить легенду на графики. Сделать выводы.

import matplotlib.pyplot as plt

import firstTask

#'age,sex,bmi,children,smoker,region,charges'

inshuranceData = firstTask.getData()

def bmiInfo(inshuranceData):

print("mean, median, mode: ", inshuranceData['bmi'].mean(),inshuranceData['bmi'].median(), inshuranceData['bmi'].mode()

plt.bar(1, inshuranceData['bmi'].mean(), color='red', label='mean')

plt.bar(2, inshuranceData['bmi'].median(), color='blue', label='median')

plt.bar(3, inshuranceData['bmi'].mode(), color='green', label='mode')

def chargesInfo(inshuranceData):

print("mean, median, mode: ", inshuranceData['charges'].mean(),inshuranceData['charges'].median(), inshuranceData['charges'].mode()

plt.bar(1, inshuranceData['charges'].mean(), color='cyan', label='mean')

plt.bar(2, inshuranceData['charges'].median(), color='magenta',label='median')

plt.bar(3, nshuranceData['charges'].mode(), color='yellow', label='mode')

bmiInfo(inshuranceData)

#chargesInfo(inshuranceData)

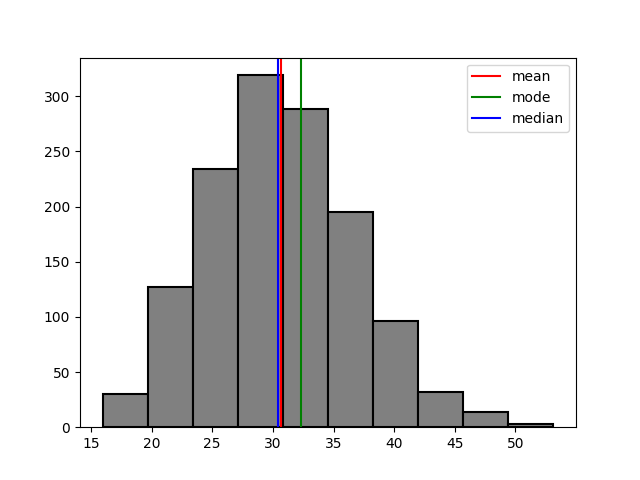
plt.legend()

plt.grid()

plt.show()

Значения и гистограмма для ‘bmi’:

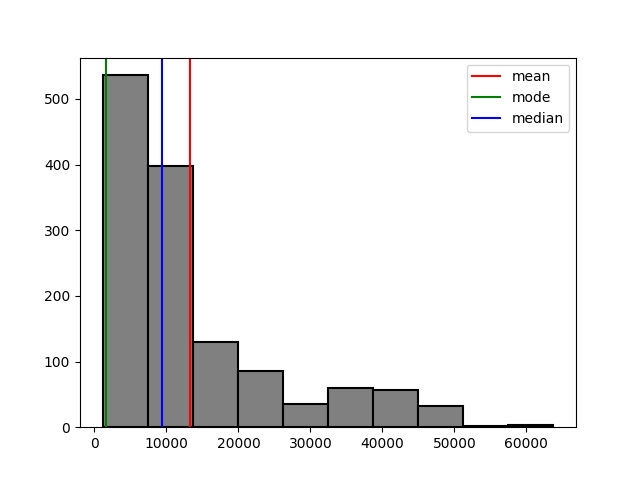




Данная гистограмма отражает разницу между средним, медианным и самым распространенным элементом датасета. Наглядно можно увидеть, что разница между первым и вторым значением не велика, однако самое распространенное значение все же выше, и ближе отображает общее содержание.

Значения и гистограмма для ‘charges’:





Данная гистограмма отображает эффект от выбросов, оказываемый на расчет среднего значения. За счет единичных показателей, выходящих за примерные рамки, в которые входят остальные показатели, существенно завышается среднее значение. Это так же может влиять на медианное значение, но уже в меньшей степени.

1. Построить box-plot для числовых показателей. Названия графиков должны соответствовать названиям признаков. Сделать выводы.

import firstTask

import matplotlib.pyplot as plt

#'age,sex,bmi,children,smoker,region,charges'

inshuranceData = firstTask.getData()

plt.figure(figsize=(5, 7))

#inshuranceData.boxplot()

#plt.boxplot(inshuranceData['age'], labels=['age'], vert=True)

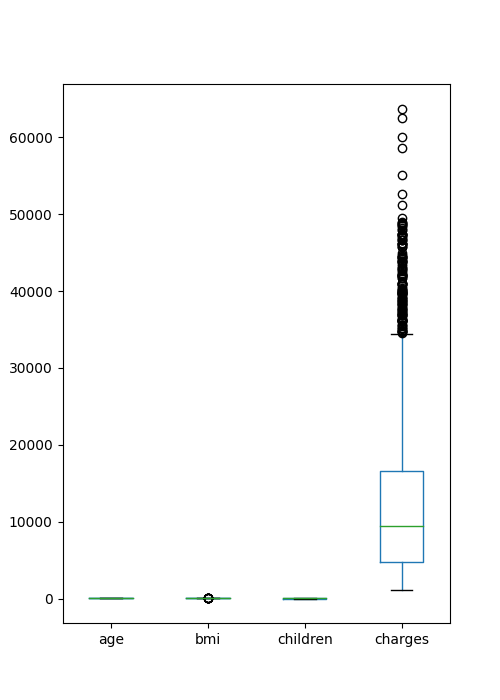
#plt.boxplot(inshuranceData['bmi'], labels=['bmi'], vert=True)

#plt.boxplot(inshuranceData['children'], labels=['children'], vert=True)

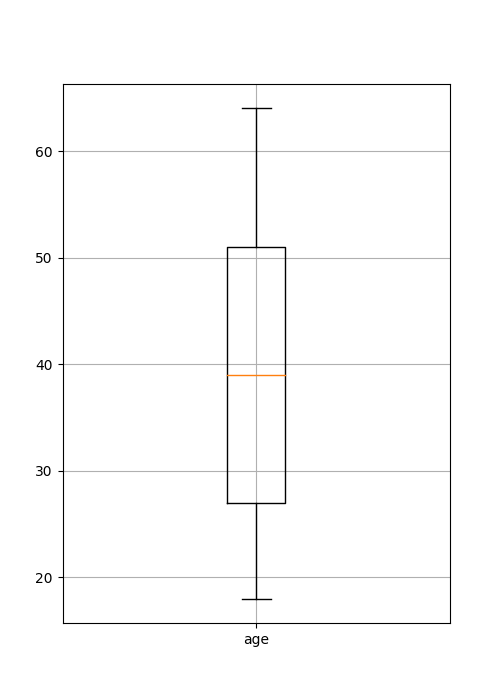
plt.boxplot(inshuranceData['charges'], labels=['charges'], vert=True)

plt.grid()

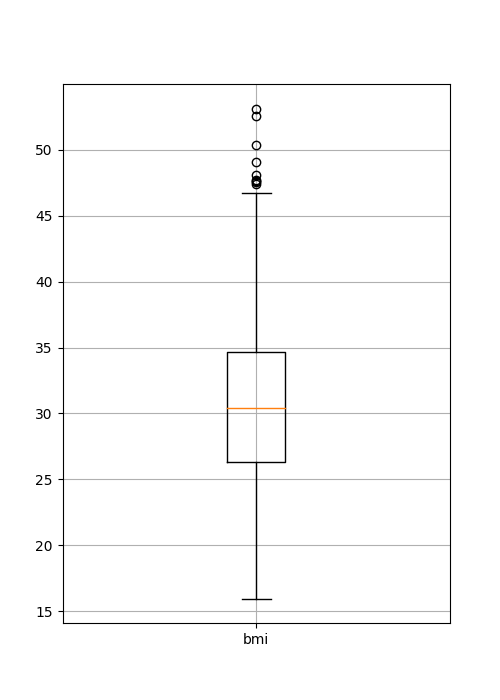
plt.show()



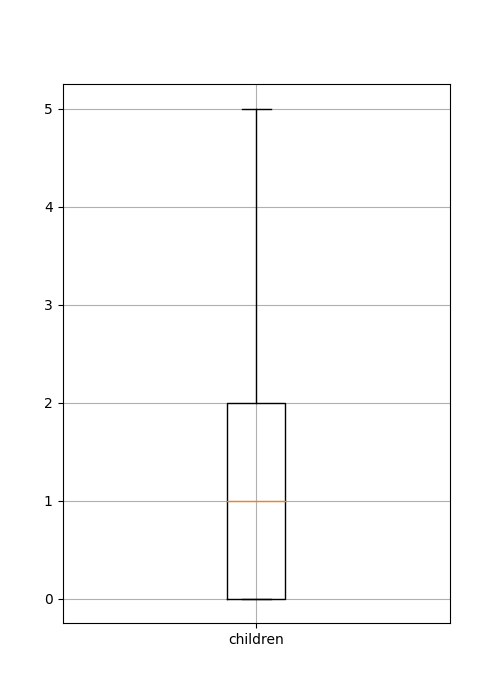
Поскольку диапазон значений у boxplot’ов очень сильно различается, было сделано несколько графиков по отдельности отображающие свои столбцы: ‘age’, ‘bmi’, ‘children’ и ‘charges’



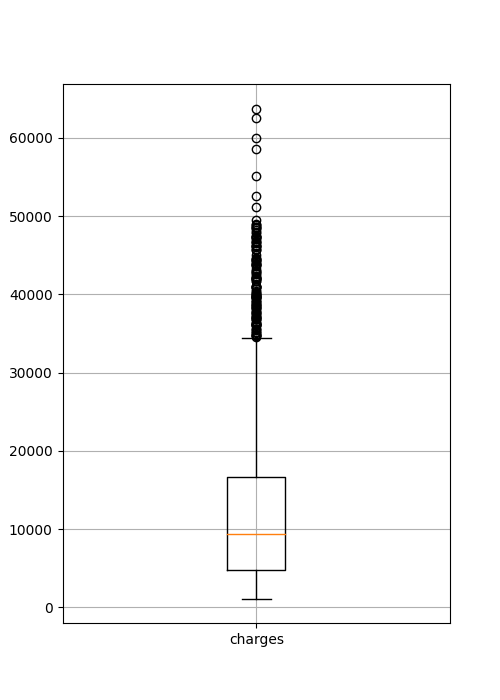
Из данного boxplot’а можно увидеть, что минимальное значение age чуть менее 20, максимальное 60 с небольшим, медианное около 39, а выборка в 50% (Q1-Q3) от всей группы лежит в диапазоне от примерно 27 до 51.



На данном boxplot’е можно увидеть, что усредненное минимальное значение bmi около 16, максимальное около 47, однако есть единичные случаи (выбросы), когда значение bmi превышает 47-48. 50% выборка (Q1-Q3) лежит в диапазоне от примерно 26 до 34.



Из данного boxplot’а можно увидеть что минимальное количество детей у людей, внесенных в данную выборку – 0 (ожидаемо), а Q1-Q3 ограничивается 2 детьми. Общее максимальное значение – 5.



Из данного boxplot’а следует, что минимальное значение суммы выплат – около 1-2 тысяч, максимальное, но при этом еще более-менее частовстречающееся – около 35 тыс., однако есть выбросы более этого значения. Q1-Q3 приходится на 5-17.5 тыс.

1. Используя признак charges или bmi, проверить, выполняется ли центральная предельная теорема. Использовать различные длины выборок n. Количество выборок = 300. Вывести результат в виде гистограмм. Найти стандартное отклонение и среднее для полученных распределений. Сделать выводы.

import firstTask

import numpy as np

import matplotlib.pyplot as plt

inshuranceData = firstTask.getData()

bmiData = inshuranceData['bmi']

means32 = []

means128 = []

means512 = []

for i in range(300):

sample = bmiData.sample(32)

means32.append(sample.mean())

sample = bmiData.sample(128)

means128.append(sample.mean())

sample = bmiData.sample(512)

means512.append(sample.mean())

fig, ax = plt.subplots(1, 3)

ax[0].hist(means32, color='cyan', label='32')

ax[0].set\_title('32')

ax[1].hist(means128, color='magenta', label='128')

ax[1].set\_title('128')

ax[2].hist(means512, color='yellow', label='512')

ax[2].set\_title('512')

print(f"среднее bmi: {bmiData.mean()}")

print("выборка в 32 элементов:")

print(f"\tстандартное отклонение: {np.std(means32)}")

print(f"\tсреднее: {np.mean(means32)}")

print("выборка в 128 элементов:")

print(f"\tстандартное отклонение: {np.std(means128)}")

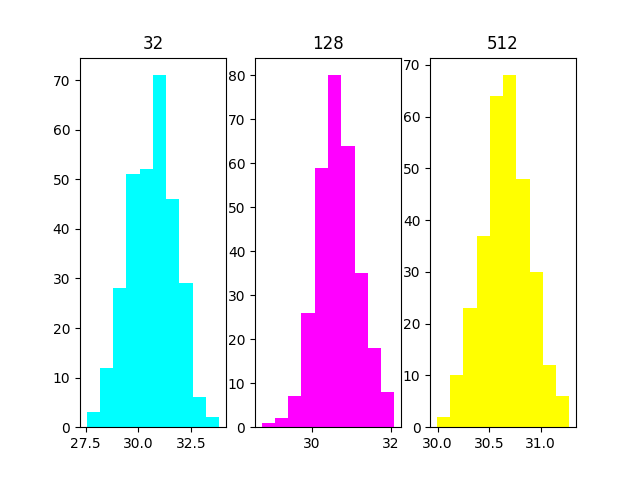
print(f"\tсреднее: {np.mean(means128)}")

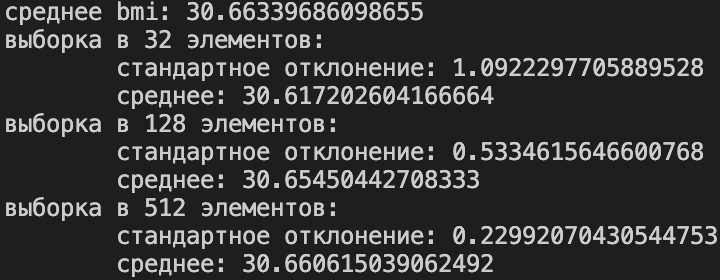
print("выборка в 512 элементов:")

print(f"\tстандартное отклонение: {np.std(means512)}")

print(f"\tсреднее: {np.mean(means512)}")

plt.show()





Выводом из данных расчетом может стать то, что при увеличении количества элементов выборки – уменьшается стандартное отклонение и незначительно растет среднее число. И при всем этом из центральной предельной теоремы следует, что чем больше элементов будет в нашей выборке, тем ближе результат будет к нормальному распределению.

1. Построить 95% и 99% доверительный интервал для среднего значения расходов и среднего значения индекса массы тела.

import firstTask

import scipy.stats as st

import numpy as np

inshuranceData = firstTask.getData()

bmi = inshuranceData['bmi']

charges = inshuranceData['charges']

bmi\_ci\_95 = st.norm.interval(0.95, loc=np.mean(bmi), scale = st.sem(bmi))

charges\_ci\_95 = st.norm.interval(0.95, loc=np.mean(charges), scale = st.sem(charges))

bmi\_ci\_99 = st.norm.interval(0.95, loc=np.mean(bmi), scale = st.sem(bmi))

charges\_ci\_99 = st.norm.interval(0.99, loc=np.mean(charges), scale = st.sem(charges))

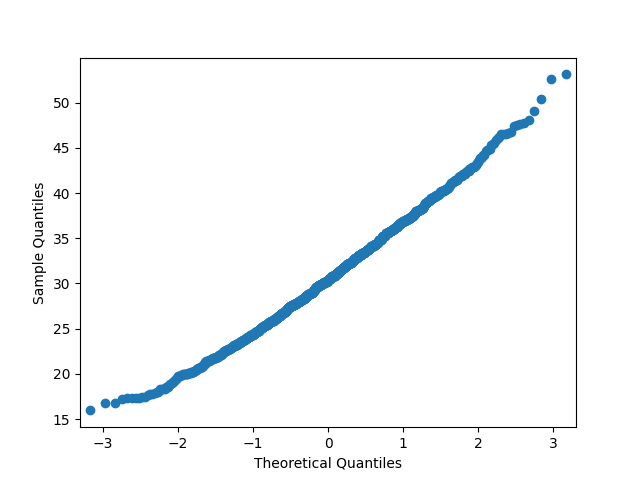
print('bmi ci 95%, 99%:', np.round(bmi\_ci\_95, 5), np.round(bmi\_ci\_99, 5))

print('charges ci 95%, 99%: ', np.round(charges\_ci\_95, 5), np.round(charges\_ci\_99, 5))

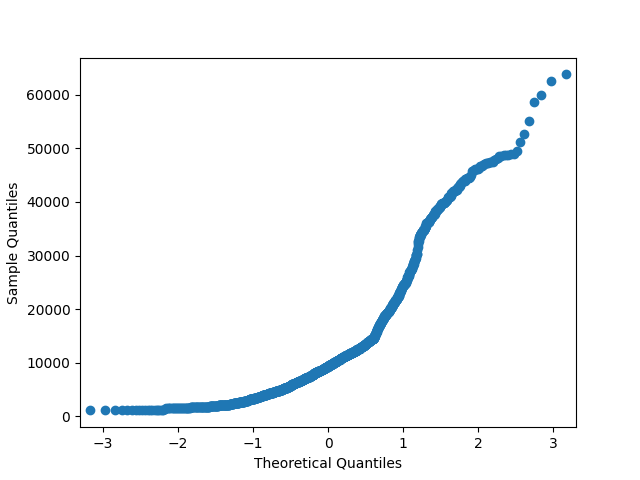


1. Проверить распределения следующих признаков на нормальность: индекс массы тела, расходы. Сформулировать нулевую и альтернативную гипотезы. Для каждого признака использовать KS-тест и q-q plot. Сделать выводы на основе полученных p-значений.

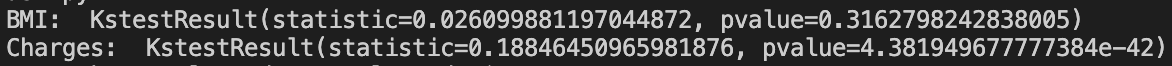
Q-q plot’ы для ‘bmi’ и ‘charges’:



По Q-Q plot’у значения bmi можно предположить, что данная выборка близка к нормальному распределению.



По Q-Q plot’у значения charges можно предположить, что данная выборка очень далека от нормального распределения.



На основе полученных данных можно сформулировать нулевую и альтернативную гипотезу для каждого значения.

Нулевая гипотеза – между группами ‘bmi’, ‘charges’ и ‘norm’ нет взаимосвязи, соответственно они не являются нормальным распределением.

Альтернативная гипотеза – эти группы являются нормальным распределением.

Исходя из Q-Q plot’ов можно сказать, что ‘bmi’ может быть близок к нормальному распределению, а вот ‘charges’ уже вряд ли. А если исходить из результатов теста Колмогорова-Смирнова, то поскольку оба p-значения больше 0.05, то обе группы (‘bmi’ и ‘charges’) не подходит под модель нормального распределения.