ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ

ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

**Московский институт электроники и математики им. А.Н. Тихонова**

Рустамова Дарина Дмитриевна, группа БИВ205,

**ОТЧЕТ**

**ПО ДОМАШНЕЙ РАБОТЕ 1**

по дисциплине «Python в науке о данных»

Тема: «Создание информационной основы для приложений»

Бригада 12

Дата сдачи отчета: 10.05.2021

Москва 2021 г.

Создание информационной основы для приложений

1. **1НФ**

Ниже указана часть таблицы из >150 записей.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Name** | **AlbumId** | **GenreId** | **Milliseconds** | **Album Title** | **ArtistId** | **Artist Name** | **Genre Name** |
| A.N.D.R.O.T.A.Z. | 14 | 3 | 266266 | Alcohol Fueled Brewtality Live! [Disc 1] | 11 | Black Label Society | Metal |
| Forgiven | 6 | 1 | 300355 | Jagged Little Pill | 4 | Alanis Morissette | Rock |
| Heaven's Dead | 11 | 4 | 276688 | Out Of Exile | 8 | Audioslave | Alternative & Punk |
| Hand In My Pocket | 6 | 1 | 221570 | Jagged Little Pill | 4 | Alanis Morissette | Rock |
| Evil Walks | 1 | 1 | 263497 | For Those About To Rock We Salute You | 1 | AC/DC | Rock |
| Intro/ Low Down | 14 | 3 | 323683 | Alcohol Fueled Brewtality Live! [Disc 1] | 11 | Black Label Society | Metal |
| Real Thing | 7 | 1 | 243879 | Facelift | 5 | Alice In Chains | Rock |
| World Of Trouble | 14 | 3 | 359157 | Alcohol Fueled Brewtality Live! [Disc 1] | 11 | Black Label Society | Metal |
| Put The Finger On You | 1 | 1 | 205662 | For Those About To Rock We Salute You | 1 | AC/DC | Rock |
| I Know Somethin (Bout You) | 7 | 1 | 261955 | Facelift | 5 | Alice In Chains | Rock |
| All For You | 14 | 3 | 235833 | Alcohol Fueled Brewtality Live! [Disc 1] | 11 | Black Label Society | Metal |
| Be Yourself | 11 | 4 | 279484 | Out Of Exile | 8 | Audioslave | Alternative & Punk |
| Blood In The Wall | 15 | 3 | 284368 | Alcohol Fueled Brewtality Live! [Disc 2] | 11 | Black Label Society | Metal |
| Doesn't Remind Me | 11 | 4 | 255869 | Out Of Exile | 8 | Audioslave | Alternative & Punk |
| 2 Minutes To Midnight | 107 | 3 | 359810 | Powerslave | 90 | Iron Maiden | Metal |
| A Estrada | 26 | 8 | 248842 | Acústico MTV [Live] | 19 | Cidade Negra | Reggae |
| A Estrada | 27 | 8 | 282174 | Cidade Negra - Hits | 19 | Cidade Negra | Reggae |
| You Oughta Know (Alternate) | 6 | 1 | 491885 | Jagged Little Pill | 4 | Alanis Morissette | Rock |
| Stratus | 13 | 2 | 582086 | The Best Of Billy Cobham | 10 | Billy Cobham | Jazz |
| Blood In The Wall | 15 | 3 | 284368 | Alcohol Fueled Brewtality Live! [Disc 2] | 11 | Black Label Society | Metal |
| Doesn't Remind Me | 11 | 4 | 255869 | Out Of Exile | 8 | Audioslave | Alternative & Punk |
| Outra Vez | 8 | 2 | 126511 | Warner 25 Anos | 6 | Antônio Carlos Jobim | Jazz |

Initial

В качестве потенциального ключа взяты атрибуты {Name, AlbumID}.

Milliseconds зависит от Name и AlbumID, потому что Name может повторяться. GenreID, ArtistID, Album Title зависят от AlbumID. Genre Name зависит от GenreID, а Artist Name от ArtistID.

1. **2НФ**

Таблицы ниже созданы на основе таблицы выше (часть полной).

Исходная таблица декомпозирована на 2: *Tracks* c атрибутами Name, AlbumId, Milliseconds и *Album\_artis\_genre* с AlbumID и остальными атрибутами.

|  |  |  |
| --- | --- | --- |
| **Name** | **AlbumId** | **Milliseconds** |
| A.N.D.R.O.T.A.Z. | 14 | 266266 |
| Forgiven | 6 | 300355 |
| Heaven's Dead | 11 | 276688 |
| Hand In My Pocket | 6 | 221570 |
| Evil Walks | 1 | 263497 |
| Intro/ Low Down | 14 | 323683 |
| Real Thing | 7 | 243879 |
| World Of Trouble | 14 | 359157 |
| Put The Finger On You | 1 | 205662 |
| I Know Somethin (Bout You) | 7 | 261955 |
| All For You | 14 | 235833 |
| Be Yourself | 11 | 279484 |
| Blood In The Wall | 15 | 284368 |
| Doesn't Remind Me | 11 | 255869 |
| 2 Minutes To Midnight | 107 | 359810 |
| A Estrada | 26 | 248842 |
| A Estrada | 27 | 282174 |
| You Oughta Know (Alternate) | 6 | 491885 |
| Stratus | 13 | 582086 |
| Blood In The Wall | 15 | 284368 |
| Doesn't Remind Me | 11 | 255869 |
| Outra Vez | 8 | 126511 |

1. Tracks

В таблицe *Album\_artist\_genre* не ключевые атрибуты зависят от части сложного ключа AlbumID.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **AlbumId** | **GenreId** | **Album Title** | **ArtistId** | **Artist Name** | **Genre Name** |
| 1 | 1 | For Those About To Rock We Salute You | 1 | AC/DC | Rock |
| 6 | 1 | Jagged Little Pill | 4 | Alanis Morissette | Rock |
| 7 | 1 | Facelift | 5 | Alice In Chains | Rock |
| 8 | 2 | Warner 25 Anos | 6 | Antônio Carlos Jobim | Jazz |
| 11 | 4 | Out Of Exile | 8 | Audioslave | Alternative & Punk |
| 13 | 2 | The Best Of Billy Cobham | 10 | Billy Cobham | Jazz |
| 14 | 3 | Alcohol Fueled Brewtality Live! [Disc 1] | 11 | Black Label Society | Metal |
| 15 | 3 | Alcohol Fueled Brewtality Live! [Disc 2] | 11 | Black Label Society | Metal |
| 26 | 8 | Acústico MTV [Live] | 19 | Cidade Negra | Reggae |
| 27 | 8 | Cidade Negra - Hits | 19 | Cidade Negra | Reggae |
| 107 | 3 | Powerslave | 90 | Iron Maiden | Metal |

2. Album\_artist\_genre

Вторая таблица не является 3НФ, потому что есть зависимость не ключевых атрибутов. Genre Name зависит от GenreID, а Artist Name зависит от ArtistID. GenreId и ArtistID зависят от AlbumId.

1. **3НФ**

Таблицы ниже основаны на полной таблице из >150 записей.

|  |  |  |  |
| --- | --- | --- | --- |
| **AlbumID** | **Title of the album** | **GenreID** | **ArtistID** |
| 1 | For Those About To Rock We Salute You | 1 | 1 |
| 2 | Balls to the Wall | 1 | 2 |
| 3 | Restless and Wild | 1 | 2 |
| 4 | Let There Be Rock | 1 | 1 |
| 5 | Big Ones | 1 | 3 |
| 6 | Jagged Little Pill | 1 | 4 |
| 7 | Facelift | 1 | 5 |
| 8 | Warner 25 Anos | 2 | 6 |
| 9 | Plays Metallica By Four Cellos | 3 | 7 |
| 10 | Audioslave | 1 | 8 |
| 11 | Out Of Exile | 4 | 8 |
| 12 | BackBeat Soundtrack | 5 | 9 |
| 13 | The Best Of Billy Cobham | 2 | 10 |
| 14 | Alcohol Fueled Brewtality Live! [Disc 1] | 3 | 11 |
| 15 | Alcohol Fueled Brewtality Live! [Disc 2] | 3 | 11 |
| 16 | Black Sabbath | 3 | 12 |
| 26 | Acústico MTV [Live] | 8 | 19 |
| 27 | Cidade Negra - Hits | 8 | 19 |
| 95 | A Real Dead One | 3 | 90 |
| 102 | Live After Death | 3 | 90 |
| 104 | Live At Donington 1992 (Disc 2) | 1 | 90 |
| 107 | Powerslave | 3 | 90 |
| 108 | Rock In Rio [CD1] | 3 | 90 |

3. Albums

|  |  |
| --- | --- |
| **ArtistID** | **Artist Name** |
| 1 | AC/DC |
| 2 | Accept |
| 3 | Aerosmith |
| 4 | Alanis Morissette |
| 5 | Alice In Chains |
| 6 | Antônio Carlos Jobim |
| 7 | Apocalyptica |
| 8 | Audioslave |
| 9 | BackBeat |
| 10 | Billy Cobham |
| 11 | Black Label Society |
| 12 | Black Sabbath |
| 19 | Cidade Negra |
| 90 | Iron Maiden |

4. Artists

|  |  |
| --- | --- |
| **GenreId** | **Genre Name** |
| 1 | Rock |
| 2 | Jazz |
| 3 | Metal |
| 4 | Alternative & Punk |
| 5 | Rock And Roll |
| 8 | Reggae |

5. Genres

**Вывод**

Таким образом, из начального отношения получилось 4 справочника: *Tracks, Artists, Genres and Albums.*

Знакомство с графическими отчетами

**Столбчатая диаграмма**

Также известна как линейчатая или полосчатая диаграмма. Классическая столбиковая диаграмма оперирует горизонтальными или вертикальными столбцами для демонстрации дискретных, числовых сравнений между разными категориями. На одной оси диаграммы представлены конкретные сравниваемые категории, а на другой – шкала дискретных значений. Дискретные данные столбиковых диаграмм – это данные по категориям, которые отвечают на вопрос: «Сколько?» - по каждой из категорий. Единственный крупный недостаток столбиковых диаграмм – это оформление условных обозначений при большом количестве полос.

Для того чтобы изобразить диаграмму в python с помощью библиотеки matplotlib, нужно использовать функцию *bar().*

*Bar() – для построения вертикальной диаграммы*

*Barh() – для построения горизонтальной диаграммы*

matplotlib.pyplot.bar(*x*, *height*, *width=0.8*, *bottom=None*, *\**, *align='center'*, *data=None*, *\*\*kwargs*)

Параметры:

X – х координата диаграммы, Height – высота диаграммы, width – ширина диаграммы(по умолчанию 0.8), bottom – у координата диаграммы, align – выравнивание по х координат.

Например, следующий код показывает столбчатую диаграмму:

import numpy as np

import matplotlib.pyplot as plt

np.random.seed(123)

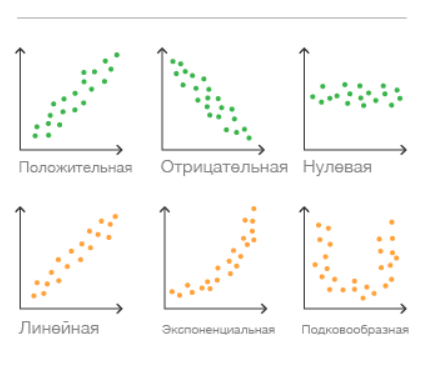
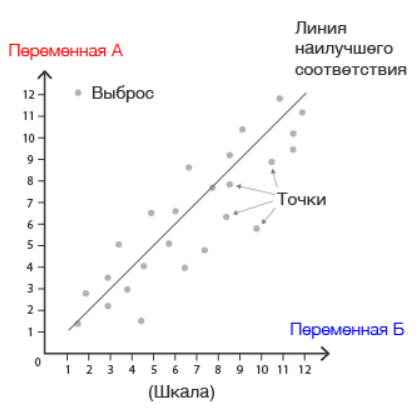
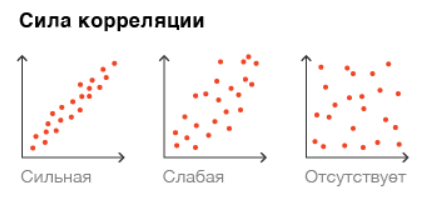
groups = ["Яблоки", "Персики", "Бананы"]

counts = np.random.randint(3, 10, len(groups))

plt.bar(groups, counts)

plt.show()

**Диаграмма рассеяния**

Также известна как график рассеяния, точечный график, координатный график или скатерограмма. На диаграммах рассеяния ряд точек, размещенных в декартовой системе координат, отображает значения по двум переменным. Присвоив каждой оси переменную, можно определить, существуют ли отношения или корреляция между этими двумя переменными. Отображаемые на диаграммах рассеяния паттерны позволяют увидеть разные типы корреляции. Среди них: **положительная** (оба значения увеличиваются), **отрицательная** (одно значение увеличивается, в то время как второе уменьшается), **нулевая** (отсутствие корреляции), **линейная**, **экспоненциальная** и **подковообразная**. Сила корреляции определяется по тому, насколько близко расположены друг от друга точки на графике. Точки, которые значительно удалены от общего кластера точек, называются **выбросами**. На графике могут использоваться линии или кривые, которые помогают при анализе и проводятся максимально близко ко всем точкам, чтобы продемонстрировать, как бы все эти точки могли потенциально выстроиться в одну линию. Этот элемент известен под названием **«линия наилучшего соответствия»** или **«линия общего направления»** и может использоваться для оценки через интерполяцию.

В библиотеке matplotlib.pyplot можно нарисовать такую диаграмму с помощью функции scatter()

matplotlib.pyplot.scatter(*x*, *y*, *s=None*, *c=None*, *marker=None*, *cmap=None*, *norm=None*, *vmin=None*, *vmax=None*, *alpha=None*, *linewidths=None*, *\**, *edgecolors=None*, *plotnonfinite=False*, *data=None*, *\*\*kwargs*)

Параметры:

X, y – данные позиций, S – ширина маркера, С – цвет, marker – стиль маркера, сmap – экземпляр цветовой карты и т.д

Пример кода на Python:

import matplotlib.pyplot as plt

import pandas as pd

df = pd.read\_csv('AmesHousing.csv')

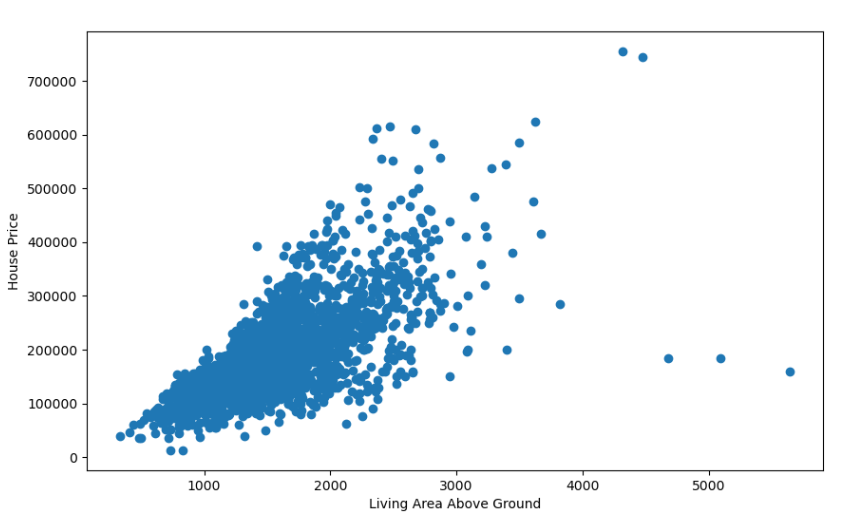
fig, ax = plt.subplots(figsize=(10, 6))

ax.scatter(x = df['Gr Liv Area'], y = df['SalePrice'])

plt.xlabel("Living Area Above Ground")

plt.ylabel("House Price")

plt.show()

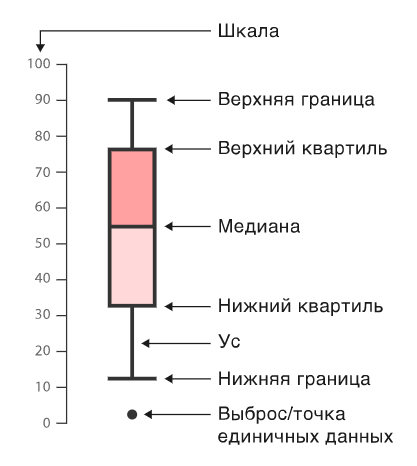


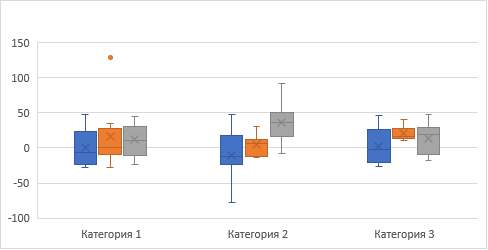
**Диаграмма Бокса-Уискера**

Диаграммы размаха («ящик с усами») (Box and Whisker Plot или Box Plot) – это удобный способ визуального представления групп числовых данных через квартили. Прямые линии, исходящие из ящика, называются «усами» и используются для обозначения степени разброса (дисперсии) за пределами верхнего и нижнего квартилей. Выбросы иногда отображаются в виде отдельных точек, находящихся на одной линии с усами. Диаграммы размаха могут располагаться как горизонтально, так и вертикально. Диаграммы размаха, как правило, используются в описательной статистике и позволяют быстро исследовать один или более наборов данных в графическом виде. Несмотря на то, что в сравнении с гистограммой или графиком плотности, этот график может показаться примитивным, его преимущество – в экономии пространства, что особенно удобно при сравнении распределений между большим количеством групп или наборов данных.

Виды наблюдений, которые можно сделать на основе ящика с усами:

* Каковы ключевые значения, например: средний показатель, медиана 25го перцентиля и так далее.
* Существуют ли выбросы и каковы их значения.
* Симметричны ли данные.
* Насколько плотно сгруппированы данные.
* Смещены ли данные и, если да, то в каком направлении.

Два из наиболее распространенных варианта ящика с усами – диаграмма размаха с переменной шириной и диаграмма размаха с метками. 



В matplotlib такая диаграмма рисуется с помощью функции boxplot.

matplotlib.pyplot.boxplot(*x*, *notch=None*, *sym=None*, *vert=None*, *whis=None*, *positions=None*, *widths=None*, *patch\_artist=None*, *bootstrap=None*, *usermedians=None*, *conf\_intervals=None*, *meanline=None*, *showmeans=None*, *showcaps=None*, *showbox=None*, *showfliers=None*, *boxprops=None*, *labels=None*, *flierprops=None*, *medianprops=None*, *meanprops=None*, *capprops=None*, *whiskerprops=None*, *manage\_ticks=True*, *autorange=False*, *zorder=None*, *\**, *data=None*)  Параметры

x – входные данные, notch – булевая переменная, которая отвечает за то рисовать ли прямоугольную диаграмму или нет, vert – если true, то рисует вертикальную диаграмму, иначе – горизонтальную, whis – положение усов, usermeduans – устанавливает значение медиан для каждого набора данных, widths – ширина ящиков (по умолчанию 0.5) и тд.

Пример программы:

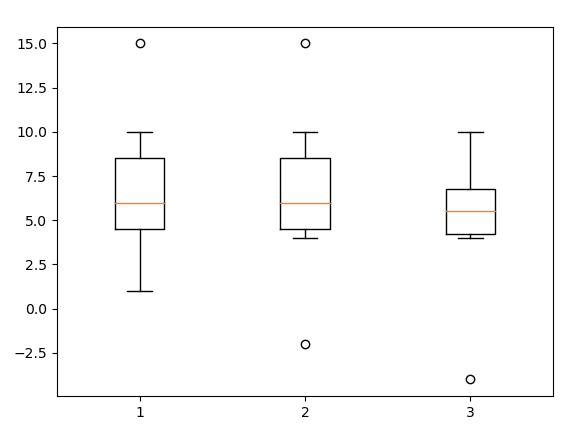
import matplotlib.pyplot as plt

plt.boxplot([[1, 5, 7, 4, 6, 10, 15],

[-2, 5, 7, 4, 6, 10, 15],

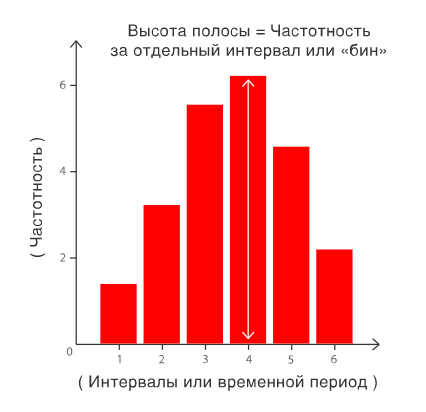
[-4, 5, 7, 4, 6, 10]])

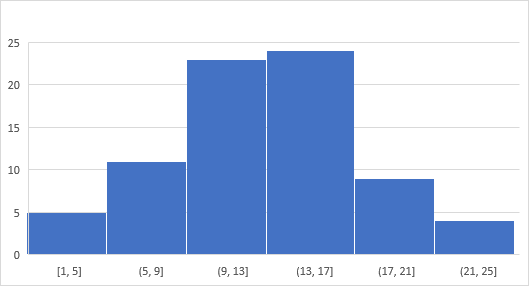
plt.show()

Оранжевая линия это медиана, круглые точки – выбросы – значения, которые находятся очень далеко от части данных, минимальное и максимальное значение это усы.

**Гистограмма**

Гистограмма визуализирует распределение данных в рамках непрерывного интервала или ограниченного периода времени. Каждая полоса на гистограмме представляет в табличной форме частотность за определенный интервал/бин. Общая площадь гистограммы равна количеству данных. Гистограммы помогают определить концентрацию значений, предельные значения и наличие пробелов или отклонений. Кроме того, они удобны для составления приблизительного обзора распределения вероятностей.





Чтобы визуализировать данные с помощью гистограммы в python, нужно воспользоваться функцией *hist().*

matplotlib.pyplot.hist(*x*, *bins=None*, *range=None*, *density=False*, *weights=None*, *cumulative=False*, *bottom=None*, *histtype='bar'*, *align='mid'*, *orientation='vertical'*, *rwidth=None*, *log=False*, *color=None*, *label=None*, *stacked=False*, *\**, *data=None*, *\*\*kwargs*)

Где:

X – входные значения(один или несколько массивов), bins – если целое, то определяет количество интервалов одинаковой ширины, если последовательность, она определяет края ячейки включая левый край первой ячейки и правый край последней ячейки; range – нижний и верхний предел bins, align – горизонтальное расположение гистограмм