ИНСТИТУТ
ЦИФРОВОГО
ОБРАЗОВАНИЯ
МГПУ

Московский городской педагогический университет
Департамент информатики, управления и технологий

Основы описательного анализа данных

МГПУ, 2023

ИНСТИТУТ
ЦИФРОВОГО
ОБРАЗОВАНИЯ
МГПУ

Основы описательного анализа данных

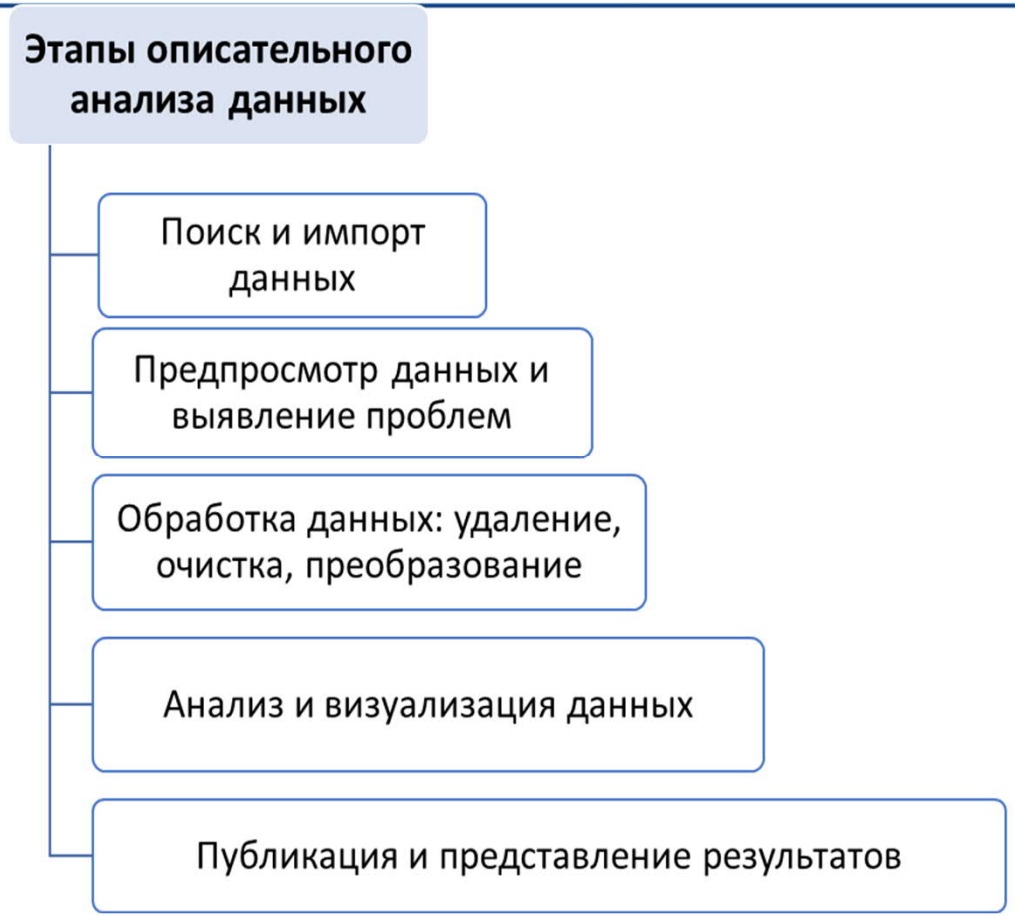
МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Общие сведения

- **В описательном анализе данных** проводится анализ данных с помощью таких методов, как
- **обобщение**,
- **агрегирование**,
- **визуализация**.
- **Рассмотрим различные типы данных**, методы их классификации, определим операции, которые можно выполнять на основе категории данных и установим процесс описательного анализа данных.

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Этапы описательного анализа данных



МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

1. Поиск и импорт данных (извлечение данных)

- Данные могут храниться в **структурированном формате** (например, базы данных или электронные таблицы) или в **неструктурированном формате** (например, веб-страницы, электронные письма, документы Word).

- После рассмотрения таких параметров, как стоимость и структура данных, нам нужно выяснить, как получить эти данные. Такие библиотеки, как **Pandas**, предоставляют функции для импорта данных в различных форматах.

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

2. Предпросмотр данных и выявление проблем

- **На этом этапе формируется представление о данных**, с которыми будет проведен анализ. Проводится отдельное изучение столбцов или объектов, значения различных сокращений и обозначений, используемых в наборе данных, представляющие собой записи или данные, а также единицы измерения, используемые для хранения данных. Необходимо задать правильные вопросы и выяснить, что необходимо сделать, прежде чем приступить к детальному анализу.

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Обработка данных

- **Этот шаг является ключевым моментом анализа данных** и самая трудоемкая работа, при этом аналитики данных тратят на это примерно **80%** своего времени. Данные в необработанном виде часто непригодны для анализа по любой из следующих причин:

- **наличие избыточных и пропущенных значений,**
- **наличие выбросов,**
- **определение некорректных типов данных,**
- **наличие посторонних данных,**
- **использование более одной единицы измерения,**
- **разброс данных,**
- **не корректная идентификация столбцов.**

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Обработка данных

- **Обработка данных** – это процесс преобразования необработанных (неструктурированных) данных таким образом, чтобы они подходили для математической обработки и построения графиков. Она включает в себя удаление или замену отсутствующих значений и неполных записей, очистку данных от пустых значений или специальных символов, таких как:

- **точки с запятой или запятые,**
- **фильтрацию данных,**
- **изменение типов данных,**
- **устранение избыточности,**
- **объединение данных с другими источниками.**

Обработка данных

При очистке данных идентифицируются переменные в наборе данных (датасет) и проводится сопоставление их со столбцами.

Последующим этапом является **структурирование данных** и контроль за тем, чтобы строки содержали наблюдения, а не признаки.

Обработка данных

Цель преобразования и подготовки данных состоит в том, чтобы иметь структуру данных, позволяющую проводить как математический, так и статистический анализ.

Анализ и визуализация данных

За этапом **обработки данных**, следует шаг – **поиск закономерностей в данных, обобщение ключевых характеристик и анализ взаимосвязей между различными функциями**.

На основе **визуализации** возможно наглядно представить важные закономерности или рассуждения в виде аналитических дашбордов. Библиотеки Python для визуализации включают следующие библиотеки:

- **Matplotlib,**
- **Seaborn,**
- **Pandas.**

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Публикация и представление результатов анализа данных

Наиболее распространенным и удобным механизмом для публикации результатов и последующего представления полученных результатов подходит использование **Блокнотов Jupyter**:

они выполняют код и служат платформой для предоставления высокоуровневой графики полученного результата.

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Структуры данных

Классификация данных

Существует два основных типа данных:

- **количественные(непрерывные),**
- **качественные(категориальные).**

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Классификация данных

Числовые значения для категориальных переменных

Категориальные данные не ограничиваются нечисловыми значениями.

Например, ранг учащегося, который может принимать такие значения, как 1/2/3 и т. д., является примером порядковой (категориальной) переменной, которая содержит числа в качестве значений. Однако эти числа не имеют математического значения; например, не имеет смысла находить среднее.

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Значение истинной нулевой точки

Интервальные переменные не имеют абсолютного нуля в качестве точки отсчета, в то время как **относительные переменные** имеют действительную нулевую точку. **Абсолютный ноль** означает отсутствие значения.

Например, такие переменные, как рост и вес, являются переменными отношения, это будет означать, что значение 0 для любой из этих переменных будет означать недопустимую или несуществующую точку данных. Для такой интервальной переменной, как температура (при измерении в градусах Цельсия или Фаренгейта), значение 0 не означает, что данные отсутствуют. 0 — это лишь одно из значений, которое может принимать переменная температуры.

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Идентификация интервальных переменных

Интервальные переменные не имеют абсолютного нуля в качестве точки отсчета, но определение переменных, обладающих этой характеристикой, может быть неочевидным.

Всякий раз, когда говорится о процентном изменении фигуры, речь, в этом случае, идет о ее предыдущем значении.

Например, процентное изменение инфляции или безработицы рассчитывается с последним значением во времени в качестве точки отсчета. Это экземпляры интервальных данных.

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Визуализация различных типов данных

При необходимости проанализировать данные, сначала определить, являются ли данные **структурированными или неструктурированными**

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Визуализация различных типов данных

В этом занятии основная задача: как классифицировать переменные в наборе данных и определить методы, которые будут применяться для каждой категории. Рассмотрим набор данных Titanic https://github.com/BosenkoTM/DAT-for-SAP/blob/main/data/titanic.csv

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Справочная информация о наборе данных

Справочная информация о наборе данных: Titanic, британское пассажирское судно, затонуло во время своего первого рейса из Саутгемптона в Нью-Йорк 15 апреля 1912 года после столкновения с айсбергом. Из 2224 пассажиров погибло 1500 человек, что сделало это событие трагедией эпических масштабов. Этот набор данных описывает статус выживания пассажиров и другие сведения о них, включая их класс, имя, возраст и количество родственников.

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Набор данных датасет titanic

	A	B	C	D	E	F	G	H	I	J	K	L
1	**PassengerId**	**Survived**	**Pclass**	**Name**	**Sex**	**Age**	**SibSp**	**Parch**	**Ticket**	**Fare**	**Cabin**	**Embarked**
2	1	0	3	Braund, M	male	22	1	0	A/5 21171	45839		S
3	2	1	1	Cumings, I	female	38	1	0	PC 17599	71.2833	C85	C
4	3	1	3	Heikkinen	female	26	0	0	STON/O2. 3101282	7.925		S
5	4	1	1	Futrelle, M	female	35	1	0	113803	53.1	C123	S
6	5	0	3	Allen, Mr.	male	35	0	0	373450	44689		S
7	6	0	3	Moran, M	male		0	0	330877	980160		Q
8	7	0	1	McCarthy,	male	54	0	0	17463	51.8625	E46	S
9	8	0	3	Palsson, M	male	2	3	1	349909	21.075		S
10	9	1	3	Johnson, M	female	27	0	2	347742	11.1333		S

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Объекты в этом наборе данных, классифицированные в соответствии с типом данных

Объекты в датасете	Описание	Уровень данных
PassengerId	идентификационный номер пассажира	номинальный
Pclass	класс пассажира (1: 1-й класс; 2: 2-й класс; 3: 3-й класс), используется в качестве показателя социально-экономического статуса пассажира.	порядковый
Survived	статус выжившего (0:не выжил; 1: выжил).	номинальный
Name	ФИО.	номинальный
Sex	Пол.	номинальный
Age	Возраст.	относительный
SibSp	количество братьев и сестер/супругов на борту.	относительный
Parch	количество родителей/детей на борту.	относительный
Ticket	номер билета.	номинальный
Fare	стоимость проезда для пассажиров (британский фунт).	относительный
Cabin	Номер каюты.	номинальный
Embarked	порт посадки (где C - Шербур, Q - Квинстаун, а S - Саутгемптон)	номинальный

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

загрузка библиотек

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
import warnings
warnings.filterwarnings("ignore")
```

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

загрузка данных в colab

```
from google.colab import files
uploaded = files.upload()
```

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

загрузка датасета в датафрейм Pandas

```
df_titanic = pd.read_csv("titanic.csv")
```

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

первичный анализ данных: типизация

```
df_titanic.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

визуальный анализ данных и выявление проблем

```
df_titanic.head()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

визуальный анализ данных и выявление проблем

```
df_titanic.describe()
```

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

визуальный анализ данных и выявление проблем

```
df_titanic.describe(include=['O'])
```

	Name	Sex	Ticket	Cabin	Embarked
count	891	891	891	204	889
unique	891	2	681	147	3
top	Braund, Mr. Owen Harris	male	347082	B96 B98	S
freq	1	577	7	4	644

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

#обработка данных: создание нового столбца данных

```
df_titanic['Cabin Letter'] = df_titanic['Cabin'].str.extract('(\w)')
df_titanic.head()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Cabin Letter
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S	NaN
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S	NaN
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S	C
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S	NaN

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

#обработка данных: замена нулевых значений на 'Unknown'

```
df_titanic['Cabin Letter'].fillna('Unknown', inplace=True)
df_titanic.head()
```

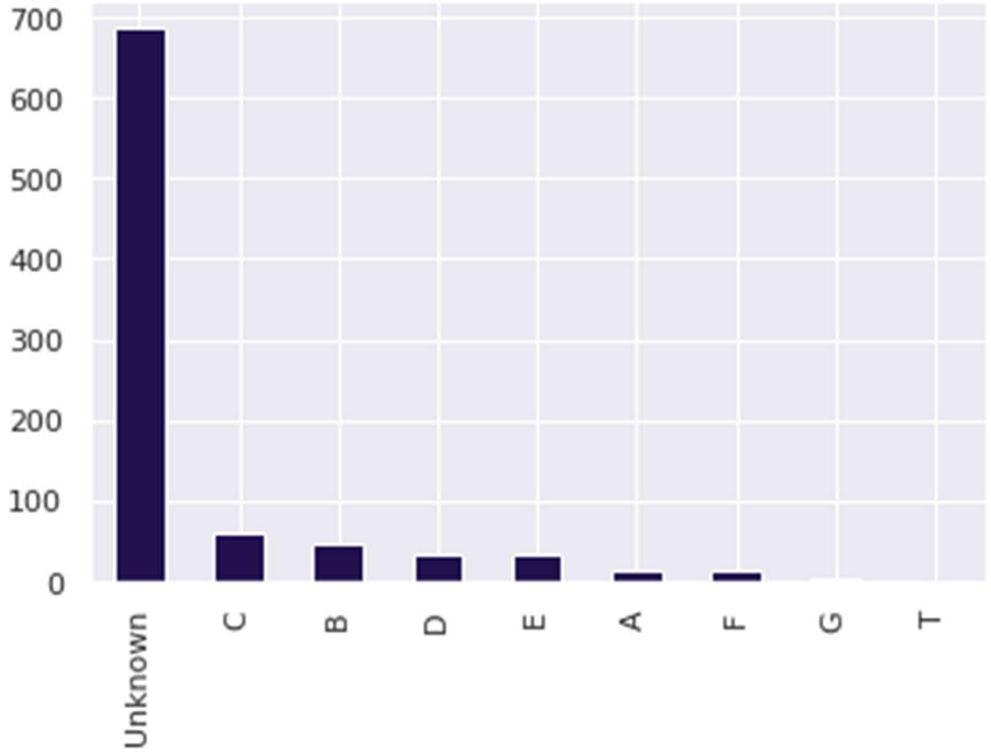
	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Cabin Letter
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S	Unknown
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S	Unknown
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S	C
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S	Unknown

#настройка публикации данных: отображения графики

```
import seaborn as sns
palette=sns.color_palette('magma')
sns.set(palette=palette)
```

визуализация гистограммы

```
df_titanic['Cabin Letter'].value_counts().plot(kind='bar')
plt.show()
```



ИНСТИТУТ
ЦИФРОВОГО
ОБРАЗОВАНИЯ
МГПУ

Обоснование классификации объектов в наборе данных titanic

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Объекты в этом наборе данных, классифицированные в соответствии с типом данных

Объекты в датасете	Описание	Уровень данных
PassengerId	идентификационный номер пассажира	номинальный
Pclass	класс пассажира (1: 1-й класс; 2: 2-й класс; 3: 3-й класс), используется в качестве показателя социально-экономического статуса пассажира.	порядковый
Survived	статус выжившего (0:не выжил; 1: выжил).	номинальный
Name	ФИО.	номинальный
Sex	Пол.	номинальный
Age	Возраст.	относительный
SibSp	количество братьев и сестер/супругов на борту.	относительный
Parch	количество родителей/детей на борту.	относительный
Ticket	номер билета.	номинальный
Fare	стоимость проезда для пассажиров (британский фунт).	относительный
Cabin	Номер каюты.	номинальный
Embarked	порт посадки (где C - Шербур, Q - Квинстаун, а S - Саутгемптон)	номинальный

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Номинальные переменные

“PassengerId”, “Survived”, “Name”, “Sex”, “Cabin” и “Embarked”

Значения неупорядоченны. Обратите внимание, что некоторые из этих переменных имеют числовые значения, но число этих значений ограничено.

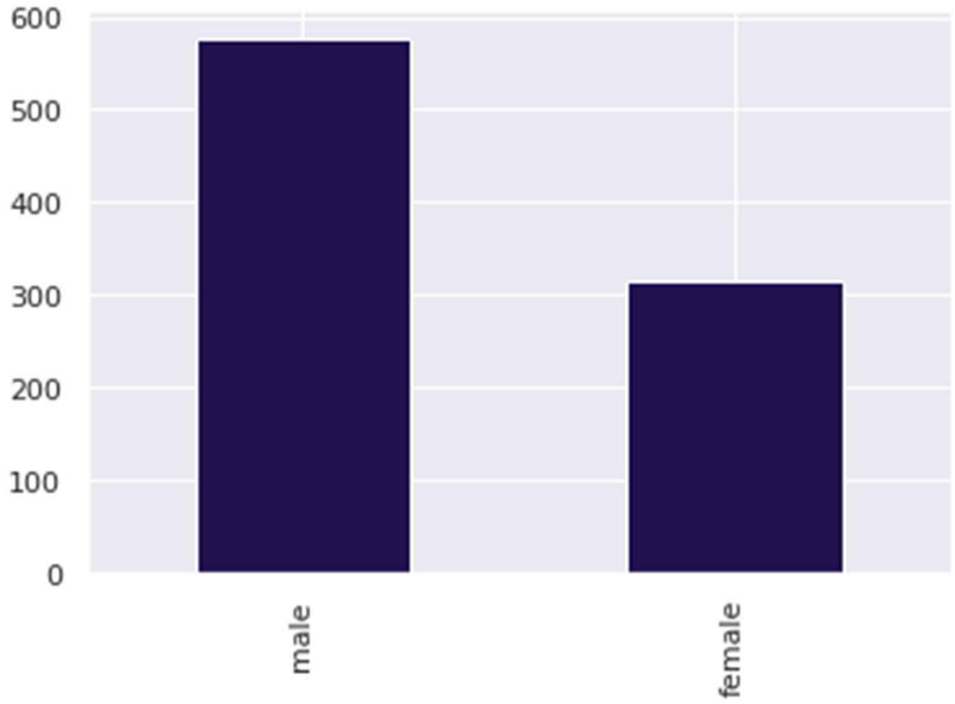
Невозможно выполнять арифметические операции с этими значениями, такие как сложение, вычитание, умножение или деление.

Одной из операций, которая является общей для номинальных переменных, является подсчет количества элементов.

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Номинальные переменные

```
df_titanic['Sex'].value_counts().plot(kind='bar')
```



МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Номинальный тип данных

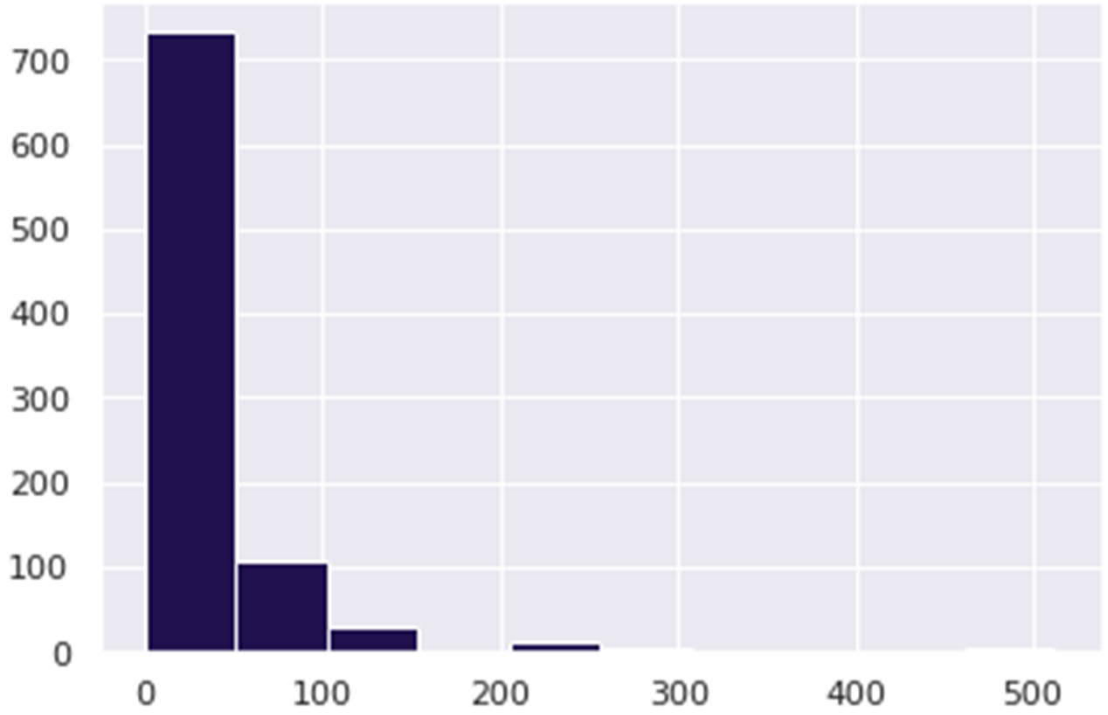
Переменные **“Age”** и **“Fare”** являются примерами данных о соотношении с нулевым значением в качестве контрольной точки.

С помощью такого типа данных выполняется широкий спектр математических операций. Например, сложение всех тарифов и деление их на общее количество пассажиров, чтобы найти среднее значение, определение стандартного отклонения. Гистограмма может быть использована для визуализации таких данных, чтобы понять распределение.

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Номинальные переменные

```
plt.hist(df_titanic['Fare'])
plt.show()
```



МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Порядковые переменные

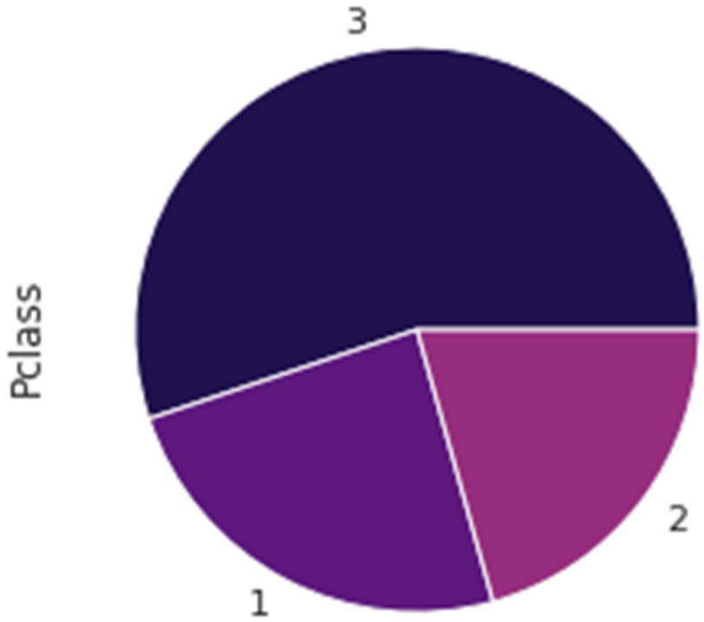
“Pclass” (или класс пассажира)

является порядковой переменной, поскольку ее значения следуют порядку. Значение 1 эквивалентно первому классу, 2 эквивалентно второму классу и так далее.

Эти классовые значения свидетельствуют о социально-экономическом статусе. Возможно узнать медианное значение и процентили, подсчитать количество значений в каждой категории, рассчитать моду и использовать графики, такие как гистограммы или круговые диаграммы.

Порядковые переменные

```
df_titanic['Pclass'].value_counts().plot(kind='pie')
plt.show()
```



ИНСТИТУТ
ЦИФРОВОГО
ОБРАЗОВАНИЯ
МГПУ

Построение графиков смешанных данных

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

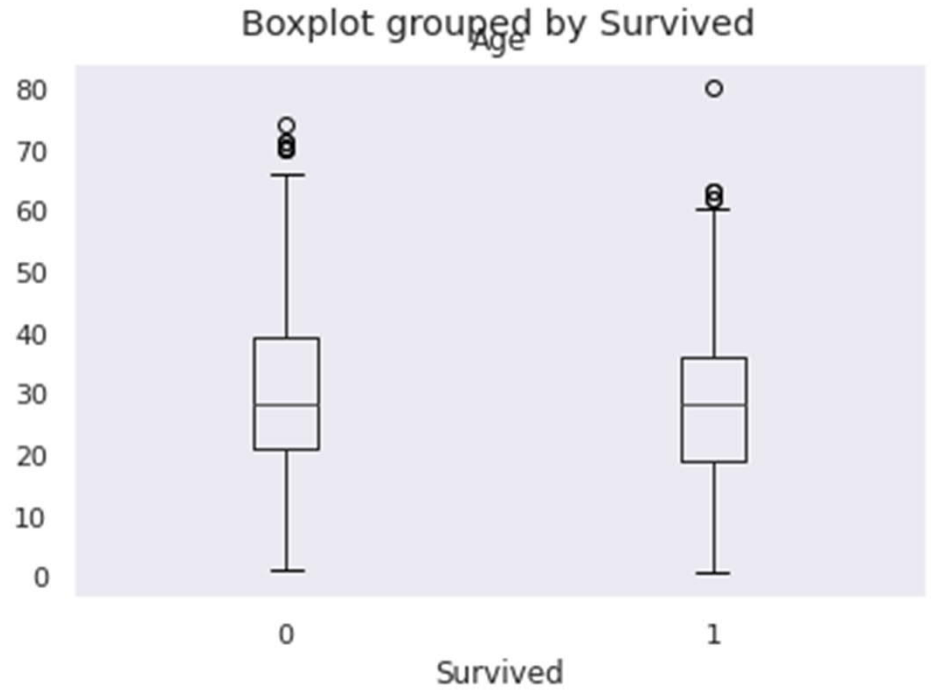
Одна категориальная и одна непрерывная переменная

Прямоугольная диаграмма(ящик с усами) показывает распределение, симметрию и выбросы для непрерывной переменной. Прямоугольная диаграмма также может отображать непрерывную переменную по отношению к категориальной переменной. Распределение **"Age"** (относительная переменная) для каждого значения номинальной переменной – **"Survived"** (0 - значение для пассажиров, которые не выжили, и 1 - значение для тех, кто выжил).

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Одна категориальная и одна непрерывная переменная

```
df_titanic.boxplot(by ='Survived', column =['Age'], grid = False)
```



МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

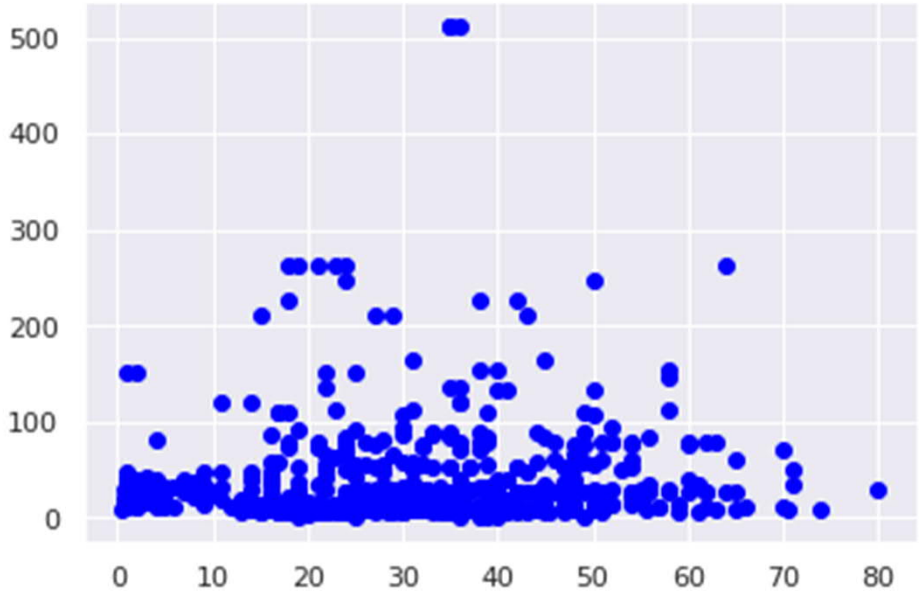
Обе непрерывные переменные

Диаграммы рассеивания используются для отображения взаимосвязи между двумя непрерывными переменными. На рисунке ниже отображаются две переменные соотношения **"Age"** и **"Fare"**, на оси x и y, чтобы получить необходимое рассеивание

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Обе непрерывные переменные

```
plt.scatter(df_titanic['Age'],df_titanic['Fare'], color='blue')
```



МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Обе категориальные переменные

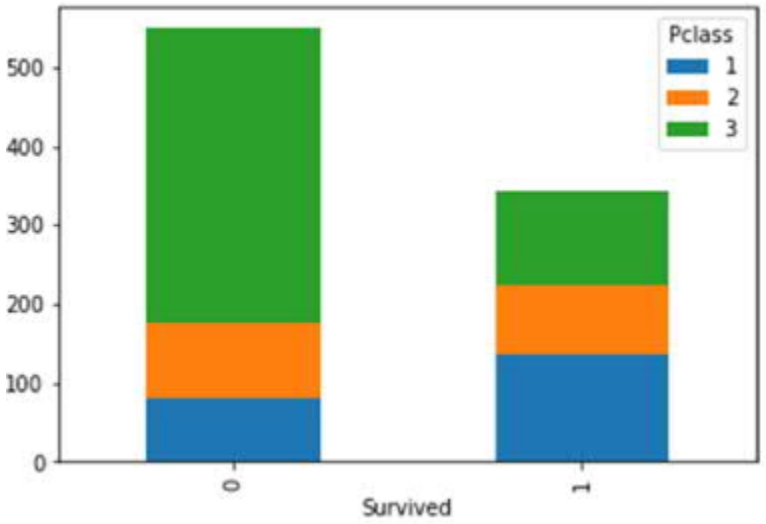
Используя кластеризованную столбчатую диаграмму, возможно объединить две категориальные переменные с изображенными рядом столбиками, чтобы представить каждую комбинацию значений для двух переменных.

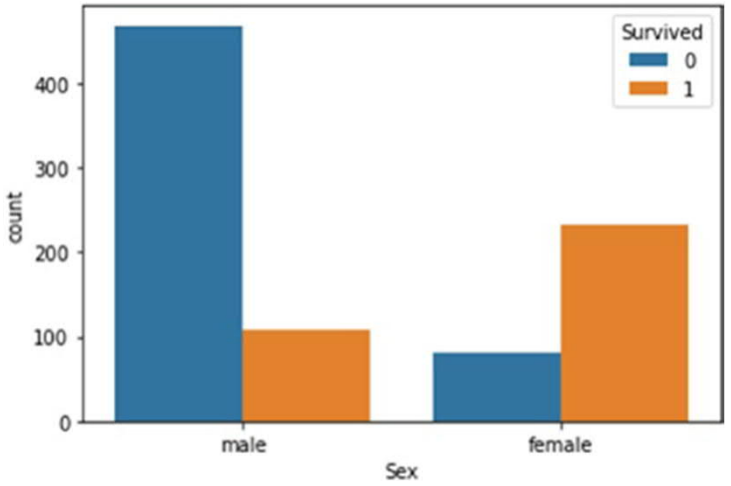
Также можно использовать стак-диаграмму для построения двух категориальных переменных. Рассмотрим следующую столбчатую диаграмму, на которой изображены две категориальные переменные – **“Pclass”** и **“Survived”.**

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Обе категориальные переменные

САМОСТОЯТЕЛЬНАЯ РАБОТА!





Описательный анализ данных представляет собой пятиэтапный процесс.

Суть процесса – обработка данных

включает в себя работу с отсутствующими значениями и другими возможными аномалиями.

ИНСТИТУТ
ЦИФРОВОГО
ОБРАЗОВАНИЯ
МГПУ

Спасибо за внимание.