

Инструменты визуализации данных с применением Python.

Слайд 2

Data science – это сбор и анализ цифровых данных, извлечение и получение информации, принятие обоснованных решений на основе этих данных и превращение их в значимые и ценные действия. Это также является междисциплинарной областью, в которой используются различные инструменты, методы и технологии, которые меняются с течением времени.

В частности, это пересечение теории вероятностей, статистики, математики, анализа данных, искусственного интеллекта, машинного обучения, информатики и бизнеса.

В контексте науки о данных существует два типа данных: традиционные и большие данные.

Традиционные данные - это данные, которые структурированы и хранятся в базах данных, которыми аналитики могут управлять с одного компьютера; они представлены в табличном формате, содержат числовые или текстовые значения.

Big Data или большие данные – это структурированные или неструктурированные массивы данных большого объема, которые поступают во все возрастающих объемах и со все большей скоростью. Обычно измеряются в терабайтах или петабайтах.

Слайд 3.

Визуализация данных – это графическое представление информации. Также это является одним из этапов процесса *data science*, который гласит, что после сбора, обработки и моделирования данных их необходимо визуализировать, чтобы можно было делать выводы.

Визуализацию можно использовать для различного объема данных. Это улучшает восприятие информации, а также ускоряет процесс анализа данных. По научным исследованиям, люди воспринимают визуальные образы лучше, чем текст, так как из всей информации поступающей в мозг изображения составляют 90%. При этом мозг обрабатывает изображения в 60 000 раз быстрее, чем текст.

Визуализация данных представляется в виде графиков, гистограмм, диаграмм, карт, пиктограмм, *2D* и *3D* – моделей.

Слайд 4.

Визуализация данных – это заключительный этап. Инструменты визуализации данных используются для демонстрации результатов анализа и стимулирования принятия решений.

Современный рынок программного обеспечения предоставляет множество инструментов визуализации данных.

К лучшим библиотекам визуализации данных, доступных на *Python* относятся: *Matplotlib*, *Plotly*, *folium* и другие.

Выбор графического отображения данных осуществляется с учетом типа данных и их предназначения. Например, *Matplotlib* – это комплексная библиотека для создания статических, анимированных и интерактивных визуализаций; *folium* визуализирует данные в виде карты-листовки; *Bokeh* используется для современных веб-браузеров, обеспечивая построение универсальной графики и высокопроизводительную интерактивность при работе с большими или потоковыми наборами данных; *Bqplot* является системой 2D визуализации для *Jupyter*, основанная на конструкциях графиков.

Слайд 5.

В качестве примера использовались данные о рейтингах регионов по сбору зерновых культур за 2020 и 2021 года. Эти данные были визуализированы в *Python* при помощи трех библиотек.

Библиотека *pandas* использовалась для возможности работы с данными из таблицы *Excel*. Кроме библиотеки *pandas* можно использовать похожие библиотеки для работы с данными: *vaes* и *polars*. Их используют для улучшения времени выполнения.

Слайд 6.

Первая библиотека *folium*

Благодаря данной библиотеке можно создать интерактивную карту, которая отображает регионы на карте и при этом показывает объём собранного зерна при помощи цвета.

На карте есть два вида разных маркеров: маркером в виде флажка отображаются регионы, входящие в рейтинг 10 за 2020 год, а маркером в виде кружочка – регионы, входящие в рейтинг 10 за 2021 год. Красным цветом обозначены регионы, в которых объём собранной продукции превышает 10 тыс. тонн; оранжевым цвет – регионы, в которых объём собранной продукции находится в диапазоне от 5 до 10 тыс. тонн; зелёным цветом – регионы, где объём собранной продукции меньше 5 тыс. тонн.

Слайд 7.

Визуализация данных при помощи библиотеки *matplotlib*. Благодаря ей были созданы две круговые диаграммы. Они обе показывают процентное соотношение количества собранных зерновых культур в рассматриваемых регионах за 2020 и 2021 года.

Слайд 8.

Последним примером является столбчатая диаграмма, также созданная при помощи библиотеки *matplotlib*. В данной программе дополнительно была использована библиотека *numpy* для создания массива чисел для оси x. Из двух таблиц были выбраны регионы, которые входили в рейтинг топ 10 регионов за 2020 и 2021 года. Это было сделано для того, чтобы сравнить, насколько вырос, или наоборот уменьшился, объем собранного зерна.

Слайд 9.

В данной исследовательской работе рассмотрены инструменты визуализации данных с применением Python. Были приведены примеры их использования. Визуализация данных является важной частью науки о данных, а также процесса их анализа, помогает человеческому мозгу воспринимать и извлекать информацию. Также позволяет оперативно визуализировать бизнес-показатели и принимать решения, основанные на данных. Было создано множество инструментов визуализации данных, которые помогают представить графически необходимую информацию различными способами и сократить время обработки информации.