UNIT-5

Data Visualization - Basic principles, ideas and tools for data visualization - Examples of inspiring (industry) projects - Exercise: create your own visualization of a complex dataset.

Exploratory Data Analysis and the Data Science Process - Basic tools (plots, graphs and summary statistics) of EDA - Philosophy of EDA - The Data Science Process - Case Study: Real Direct (online real estate firm).

Data Visualization:

Data visualization is the visual representation of the data. With the help of pictures, charts, maps, and other graphical elements, these tools provide a simple and coherent way to clearly see and easily discover insights and patterns in the data. It is a way to communicate complex information in a visual and intuitive manner, making it easier for people to understand and analyze the data. By transforming raw data into visual representations, data visualization allows patterns, trends, and insights to be easily identified and interpreted.

principles of data visualization:

Clarity and Simplicity: Ensure that the visualization is clear and easy to understand. Avoid unnecessary complexity and confusion in the visual representation.

Purposeful Design: Clearly define the purpose of the visualization. Know what message or insight you want to convey. Choose the appropriate type of visualization (e.g., bar chart, line graph, pie chart) based on the nature of the data and the message you want to communicate.

Data Accuracy: Ensure the accuracy of the data being presented. Any inaccuracies can lead to misleading interpretations. Clearly label axes, provide units, and include icons if needed.

Consistency: Maintain consistency in the use of colors, symbols, and scales across the visualization. Consistent use of elements makes it easier for the audience to interpret the information.

Intuitive Interpretation: Design visualizations in a way that allows for intuitive interpretation. Users should be able to understand the information without a detailed explanation. Use familiar symbols and conventions to enhance understanding.

Effective Use of Colour: Choose a colour that is visually attracting and aids in conveying the message.

Proper Labelling: Label all axes, data points, and any other relevant elements in the visualization. Include a title that clearly communicates the main message of the visualization.

Appropriate Scale: Choose appropriate scales for axes to avoid distorting the representation of data. Consider logarithmic scales if the data spans multiple orders of magnitude.

Interactivity (if applicable): If using interactive visualizations, ensure that the interactivity adds value and doesn't distract from the main message. Allow users to explore the data in a meaningful way.

Report: Arrange the elements of the visualization in a logical sequence to convey a narrative. Guide the viewer's attention to the most important aspects of the data.

https://www.owox.com/blog/articles/data-visualization/
ideas https://uxdesign.cc/20-ideas-for-better-data-visualization-73f7e3c2782d

Uses of Data Visualization:

- It helps to identify trends and patterns in data.
- It effectively convey the complex information.
- It facilitates the informed decision making.
- It discovers insights and relationships in data.

- It monitors performance and changes over time.

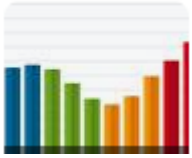
Different types of graphs used in data visualization:

Chart:

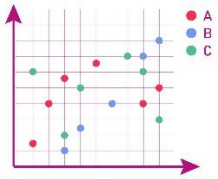
A chart is a graphical representation for data visualization, in which "the data is represented by symbols, such as bars in a bar chart, lines in a line chart, or slices in a pie chart".

A pie chart is a circular statistical graphic which is divided into slices to illustrate numerical proportion. In a pie chart, the arc length of each slice is proportional to the quantity it represents.

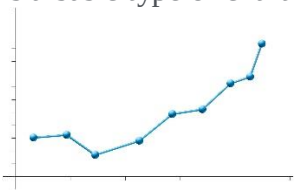
A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally. A vertical bar chart is sometimes called a column chart.



A scatter plot is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data. If the points are coded, one additional variable can be displayed.

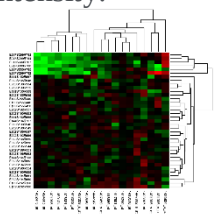


A line chart or line graph, also known as curve chart, is a type of chart which displays information as a series of data points called 'markers' connected by straight line segments. It is a basic type of chart common in many fields.

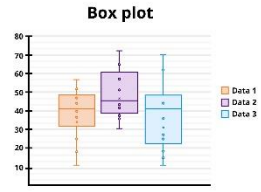


A histogram is an approximate representation of the distribution of numerical data. The term was first introduced by Karl Pearson.

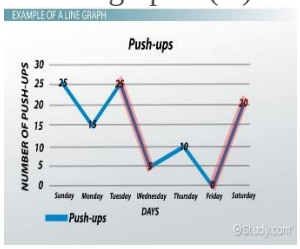
A heat map is a 2-dimensional data visualization technique that represents the magnitude of individual values within a dataset as a color. The variation in color may be by hue or intensity.



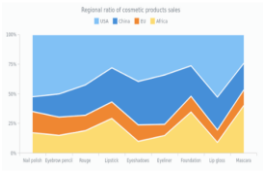
In descriptive statistics, a box plot or boxplot is a method for graphically demonstrating the locality, spread and skewness groups of numerical data through their quartiles.



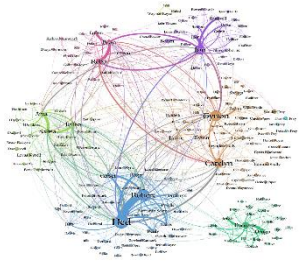
In the mathematical discipline of graph theory, the line graph of an undirected graph G is another graph L(G) that represents the adjacencies between edges of G.



Area charts, similar to line charts, are also used for tracking data over time. However, in an area chart, the space between the plotted line and the x-axis is shaded or colored for visibility. This is particularly useful for highlighting the difference between multiple variables, or for measuring overall volumes (rather than highlighting the difference between discrete data points).



network graphs, which are used to show how different elements of a network relate to one another. Each element in a network graph is represented by an individual node, interconnected to related nodes via lines. This approach is excellent for visualizing clusters within the larger whole.



Ideas of data visualization:

1. Choose the right chart type.
2. Use correct plotting directions based on the positive and negative values .
3. Always start a bar chart at 0 baseline
4. Use adaptive y-axis scale for line charts.
5. Consider your time series when using a line chart.
6. Do not use “smoothed” line charts.
7. Avoid confusing dual-axis.
8. Limit the number of slices displayed in a pie chart.
9. Label directly on the chart
10. Don’t label on top of slices.
11. Order pie slices for faster scanning
12. Avoid randomness.
13. Thin donut charts are impossible to read.
14. Let data speak for itself.
15. Pick a color palette that matches the nature of your data.
16. Design for accessibility
17. Focus on readability.
18. Use a horizontal bar chart instead of rotating labels.
19. Choose your charting library.
20. Go beyond static reports.

Data Visualization Tools:

Data Visualization Tools are software platforms that provide information in a visual format such as a graph, chart, etc to make it easily understandable and usable. Data Visualization tools are so popular as they allow analysts and statisticians to create visual data models easily according to their specifications by conveniently providing an interface, database connections, and Machine Learning tools all in one place.

1. Tableau

Tableau: Tableau is very famous as it can take in data and produce the required data visualization output in a very short time. And it can do this while providing the highest level of security with a guarantee to handle security issues as soon as they arise or are found by users.

Tableau also allows its users to prepare, clean, and format their data and then create data visualizations to obtain actionable insights that can be shared with other users.

2. Looker

Looker is a data visualization tool that can go in-depth into the data and analyze it to obtain useful insights. It provides real-time dashboards of the data for more in-depth analysis. Looker data visualizations can be shared with anyone using any particular tool. Also, you can

export these files in any format immediately. It also provides customer support wherein you can ask any question and it shall be answered.

3. Zoho Analytics

Zoho Analytics is a Business Intelligence and Data Analytics software, You can obtain data from multiple sources and mesh it together to create multidimensional data visualizations that allow you to view your business data across departments. In case you have any questions, you can use Zia which is a smart assistant created using artificial intelligence, machine learning, and natural language processing.

Zoho Analytics allows you to share or publish your reports with your colleagues and add comments or engage in conversations as required. You can export Zoho Analytics files in any format such as Spreadsheet, MS Word, Excel, PPT, PDF, etc.

4. Sisense

Sisense is a business intelligence-based data visualization system and it provides various tools that allow data analysts to simplify complex data and obtain insights for their organization and outsiders.

It is very easy to set up and learn Sisense. It can be easily installed within a minute and data analysts can get their work done and obtain results instantly. Sisense also allows its users to export their files in multiple formats such as PPT, Excel, MS Word, PDF, etc. Sisense also provides full-time customer support services whenever users face any issues.

5. IBM Cognos Analytics

IBM Cognos Analytics is an Artificial Intelligence-based business intelligence platform that supports data analytics among other things. You can visualize as well as analyze your data and share actionable insights with anyone in your organization. Even if you have limited or no knowledge about data analytics, you can use IBM Cognos Analytics easily as it interprets the data for you and presents you with actionable insights in plain language.

6. Qlik Sense

Qlik Sense is a data visualization platform that helps companies to become data-driven enterprises by providing an associative data analytics engine, sophisticated Artificial Intelligence system, and scalable multi-cloud architecture that allows you to deploy any combination of SaaS, on-premises, or a private cloud.

7. Domo

Domo is a business intelligence model that contains multiple data visualization tools that provide a consolidated platform where you can perform data analysis and then create interactive data visualizations that allow other people to easily understand your data conclusions.

8. Microsoft Power BI

Microsoft Power BI is a Data Visualization platform focused on creating a data-driven business intelligence culture in all companies today. To fulfill this, it offers self-service analytics tools that can be used to analyze, aggregate, and share data in a meaningful fashion.

9. Klipfolio

Klipfolio is a Canadian business intelligence company that provides one of the best data visualization tools. You can access your data from hundreds of different data sources like spreadsheets, databases, files, and web services applications by using connectors. Klipfolio also allows you to create custom drag-and-drop data

10. SAP Analytics Cloud

SAP Analytics Cloud uses business intelligence and data analytics capabilities to help you evaluate your data and create visualizations in order to predict business outcomes. It also provides you with the latest modeling tools that help you by alerting you of possible errors in the data and categorizing different data measures and dimensions. SAP Analytics Cloud also suggests Smart Transformations to the data that lead to enhanced visualizations.

11. Yellowfin

Yellowfin is a worldwide famous analytics and business software vendor that has a well-suited automation product that is specially created for people who have to take decisions within a short period of time.

12. Whatagraph

Whatagraph is a seamless integration that provides marketing agencies with an easy and useful way of sharing or sending marketing campaign data with clients.

13. Dundas BI

Dundas Bi is a flexible business intelligence and analysis tool. One can create and display animated dashboards, reports or scorecards. This platform can be used for data analysis can be used flexibly, openly and completely configurable.

A Sample of Data Visualization Projects:

Exploratory data analysis:

It is a method used to analyze and summarize data sets.

Exploratory data analysis (EDA) is used to analyse and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers we need, making it easier to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

EDA is primarily used to see what data can reveal beyond the formal modeling or hypothesis testing task and provides a better understanding of data set variables and the relationships between them. It can also help determine if the statistical techniques that are considering for data analysis are appropriate.

The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

Data scientists can use exploratory analysis to ensure the results they produce are valid and applicable to any desired business outcomes and goals. EDA can help answer questions about standard deviations, categorical variables, and confidence intervals. Once EDA is complete and insights are drawn, its features can then be used for more sophisticated data analysis or modeling, including machine learning.

Steps Involved in Exploratory Data Analysis (EDA)

- Data Collection
- Finding all Variables and Understanding Them
- Cleaning the Dataset
- Identify Correlated Variables
- Choosing the Right Statistical Methods
- Visualizing and Analyzing Results

EDA tools(Plots, Graphs, Summary statestics):
Specific statistical functions and techniques that can perform with EDA tools include:

- Clustering and dimension reduction techniques, which help create graphical displays of high-dimensional data containing many variables.
- Univariate visualization of each field in the raw dataset, with summary statistics.
- Bivariate visualizations and summary statistics that allows to assess the relationship between each variable in the dataset and the target variable looking at.
- Multivariate visualizations, for mapping and understanding interactions between different fields in the data.
- K-means Clustering is a clustering method in unsupervised learning where data points are assigned into K groups, i.e. the number of clusters, based on the distance from each group's centroid. The data points closest to a particular centroid will be clustered under the same category. K-means Clustering is commonly used in market segmentation, pattern recognition, and image compression.
- Predictive models, such as linear regression, use statistics and data to predict outcomes.

Some common types of multivariate graphics include:

A) Scatter Plot

The essential graphical EDA technique for two quantitative variables is the scatter plot, so one variable appears on the x-axis and the other on the y-axis and, therefore, the point for every case in your dataset. This can be used for bivariate analysis.

B) Multivariate Chart

A Multivariate chart is a type of control chart used to monitor two or more interrelated process variables. This is beneficial in situations such as process control, where engineers are likely to benefit from using multivariate charts. These charts allow monitoring multiple parameters together in a single chart. A notable advantage of using multivariate charts is that they help minimize the total number of control charts for organizational processes. Pair plots generated using the Seaborn library are a good example of multivariate charts as they help visualize the relationships between all numerical variables in the entire dataset at once.

C) Run Chart

A **run chart** is a data line chart drawn over time. In other words, a run chart visually illustrates the process performance or data values in a time sequence. Rather than summary statistics, seeing data across time yields a more accurate conclusion. A trend chart or time series plot is another name for a run chart. The plot below depicts dummy values of sales over a period of time.

D) Bubble Chart

Bubble charts scatter plots that display multiple circles (bubbles) in a two-dimensional plot. These are used to assess the relationships between three or more numeric variables. In a bubble chart, every single dot corresponds to one data point, and the values of the variables for each point are indicated by different positions such as horizontal, vertical, dot size, and dot colors.

E) Heat Map

A heat map is a colored graphical representation of multivariate data structured as a matrix of columns and rows. The heat map transforms the correlation matrix into color coding and represents these coefficients to visualize the strength of correlation among

variables. It assists in finding the best features suitable for building accurate Machine Learning models.
Apart from the above, there is also the 'Classification or Clustering analysis' technique used in EDA. It is an unsupervised type of machine learning used for the classification of input data into specified categories or clusters exhibiting similar characteristics in various groups. This can be further used to draw important interpretations in EDA.

Summary Statistics of Exploratory Data Analysis:

https://www.hcbravo.org/IntroDataSci/bookdown-notes/exploratory-data-analysis-summary-statistics.html

Exploratory Data Analysis (EDA) involves analyzing and summarizing the main characteristics of a dataset to gain insights and identify patterns. Summary statistics play a crucial role in EDA as they provide a concise overview of the dataset. Here are some common summary statistics used in exploratory data analysis:

Central Tendency:

- Mean: The average of the values in the dataset.
- Median: The middle value of the dataset when it is sorted in ascending order.
- Mode: The most frequently occurring value in the dataset.

Variability or Dispersion:

- Range: The difference between the maximum and minimum values in the dataset.
- Variance: The average of the squared differences from the mean.
- Standard Deviation: The square root of the variance, providing a measure of the spread of the data.

Shape of the Distribution:

- Skewness: A measure of the asymmetry of the distribution.
- Kurtosis: A measure of the "tailedness" of the distribution.

Quantiles:

- Quartiles: Values that divide the dataset into four equal parts.
- Percentiles: Values that divide the dataset into hundred equal parts.

Frequency Distribution:

- Histogram: A graphical representation of the distribution of a dataset.
- Frequency Tables: Tabular representation of the number of occurrences of each unique value in the dataset.

Outliers:

- Identification of Outliers: Observations that significantly differ from the rest of the dataset.

Correlation:

- Correlation Coefficients: Measures the strength and direction of linear relationships between two variables.
- Correlation Matrix: Displays the correlation coefficients between multiple variables.

Missing Values:

- Count of Missing Values: Number of observations with missing data.
- Imputation Strategies: Methods used to handle missing values.

Data Distributions:

- Box Plots: Visual representation of the distribution of a dataset.
- Probability Plots: Graphical tools to assess if a dataset follows a particular distribution.

Cross-Tabulation:

- Contingency Tables: Used to show the distribution of one variable relative to another.

These summary statistics help analysts and data scientists understand the overall characteristics of the dataset, identify potential issues, and guide further analysis. EDA is an iterative process, and these statistics are often complemented by data visualization techniques to provide a comprehensive understanding of the data.

The philosophy of Exploratory Data Analysis:

EDA is a broader approach to analyzing data. Instead of assuming a specific model for the data, EDA lets the data itself reveal its patterns and structure. EDA is a set of techniques; it's a philosophy guiding how we explore a dataset, what we look for, how we look, and how we interpret the findings. While EDA relies heavily on statistical graphics techniques, but it's not identical to statistical graphics alone. The philosophy of Exploratory Data Analysis (EDA) is like being a investigator for data. Instead of starting with a fixed idea about what the data should look like, EDA encourages us to be open-minded and let the data reveal its secrets.

The basic idea is:
Letting Data Speak: Instead of telling the data what to say, we listen to what it's trying to tell us. We don't make strong assumptions straightforward.
Seeing the Big Picture: EDA wants us to look at the whole picture, not just parts of it. It's like looking at a puzzle without knowing what the final picture is supposed to be.
Using Pictures to Understand: Instead of moving around+ numbers, we use charts and graphs to visualize the data. It's easier to see patterns and trends this way.
Being Flexible: There's no strict rulebook. EDA lets us use different tools and techniques based on what makes sense for the specific data we're exploring.
Finding Surprises: We're like detectives looking for unexpected things in the data—stuff that might be interesting or important. This includes spotting outliers or unusual patterns.
Learning as You Go: EDA is not a one-time thing. We keep going back and forth, refining our understanding each time. It's an ongoing learning process.

In simple terms, EDA is about exploring data with an open mind, using pictures to understand it, and being like a detective to uncover interesting stories the data might be hiding.

THE DATA SCIENCE PROCESS - CASE STUDY, REAL DIRECT (ONLINE REALESTATE)(consider last one)

Data Science in Pharmaceutical Industries:

Through improved data processing and cloud variety of patient information datasets. In the pharmaceutical industry, artificial intelligence and data analytics have revolutionized oncology. With new pha it is difficult for physicians to keep up into a highly competitive market for more standardized medical care options. However, with the advances in technology and the development of parallel pipelined computational models, it is now easier for the pharmaceutical industry to have a competitive advantage over the market. With various statistical models such as Markov Chains, it is now possible to predic that doctors will prescribe medicines based on their experience with the brand. In the same way, improving learning is beginning to develop itself in the area of digital marketing. It is used to understand the patterns of digital participation of physicians and their prescriptions. The main aim of this case study of data science is to discuss the problems facing it and how data science offers solutions to them..

Predictive Modeling for Maintaining Oil and Gas Supply :

The crude oil and gas industries are faced with a major problem of equipment failures, typically due to the inefficiency of the oil wells and their output at a subpar stage. With the implementation of a effective strategy that advocates for predictive maintenance, well operators can be alerted, as well as informed of maintenance times, to the critical phases of shutdown. This would lead to an increase in oil production and avoid further losses. Data Scientists can apply the Predictive Maintenance Strategy to the use of data to optimize high value machinery for the production and refining of oil products. With telemetry data extracted through sensors, a steady stream of historical data can be used to train our machine learning model. This machine learning model will predict the failure of machine parts and will alert operators of timely maintenance in order to avoid oil losses. The Data Scientist assigned to the implementation of the PdM strategy should help prevent hazards and predict machine failure, encouraging operators to take precautions.

Data Science in BioTech :

The human genome consists of four building blocks – A, T, C and G. Our appearance and characteristics are determined by the three billion permutations of these four building blocks. Although there are genetic defects and lifestyle defects, the results will lead to chronic diseases. Identifying these defects at an early stage will allow doctors and testing teams to take preventive action. Helix is one of the companies for genome analysis that provides customers with their genomic data. Also, due to the emergence of new computational methodologies, many medicines adapted to particular genetic designs have become increasingly popular. Thanks to the data explosion, we can understand and analyze complex genomic sequences on a wide scale. Data Scientists can use modern computational resources to manage massive databases and understand patterns in genomic sequences to detect defects and provide information to physicians and researchers. In addition, with the use of wearable tools, data scientists may use the relationship between genetic characteristics and medical visits to build a predictive modeling framework.

Data Science in Education:

Data Science has also changed the way students communicate with teachers and assess their success. Instructors may use data science to evaluate the input they obtain from students and use it to enhance their teaching. Data Science can be used to construct predictive modeling that can predict student drop-out levels based on their results and advise instructors to take the appropriate precautions.

Real estate:

RAW DATA COLLECTION AND PROCESSING

Since data for all the identified features was not available from a single source, we developed an automated data ingestion ecosystem to receive data from multiple public/private and free/paid sources APIs were configured to collect feature data, based on the frequency of its

publishing/updation. This data was stored on our Big Data environment To process such a large volume of data we put in place a Big Data based solution using Map/Reduce, Hadoop, R and Python systems to crunch very large volumes of data growing on a daily basis.

EXPLORATORY ANALYSIS AND FEATURE ENGINEERING

Using the cleaned and prepared data derived from data processing, our data scientists worked on readying the data for use by the algorithms that were to be built The team brainstormed to come up with Factors that drive real estate investment decisions Features essential to the statistical model Exploratory analysis and feature engineering techniques were used to reduce the number of features to only the most integral ones, that would produce the most accurate and relevant insights when modelled.

BUILDING MACHINE LEARNING ALGORITHMS AND STATISTICAL MODELS

Taking the parameters selected from the feature engineering step, we used open source tools/programming languages R, Python and Hadoop, to build machine learning algorithms CASE STUDY : HOMEUNION Rationalizing Real Estate Investment Decisions Using Data Science Various regression and machine learning techniques were used to build models for: Neighborhood Investment Rating: Rating every US neighborhood for its investment potential REALestimate: Forecasting the return on investment on every property Predicting the right offer for anchoring an optimal winning bid Time-series based Price trends on various US geographies Predicting the likely rent for every property in the US.

DATA DRIVEN INSIGHTS AND ANALYSIS

The machine learning algorithms and statistical models were run using an automated flow built on the Big Data environement CASE STUDY : HOMEUNION Rationalizing Real Estate Investment Decisions Using Data Science The relevant feature data was then modeled to generate trends/forecasts for investment risk, sale price range, rent prediction and price appreciation, etc. at multiple geographic levels