Chapter01IntroductionToHealthDataScience

January 11, 2024

1 1 | INTRODUCTION TO HEALTH DATA SCIENCE

Note: This is a Jupyter Notebook. If you are not familiar with Jupyter Notebooks, please read the Jupyter Notebook Quick Start Guide.

1.1 Python packages used in this notebook

This course uses the Python computer language and Jupyter notebooks for the generation of the course content. At the start of each notebook, we will list the Python packages used in that notebook. The following Python packages are imported for use in this notebook.

[1]:
```
from pandas import read_csv
import pandas
```

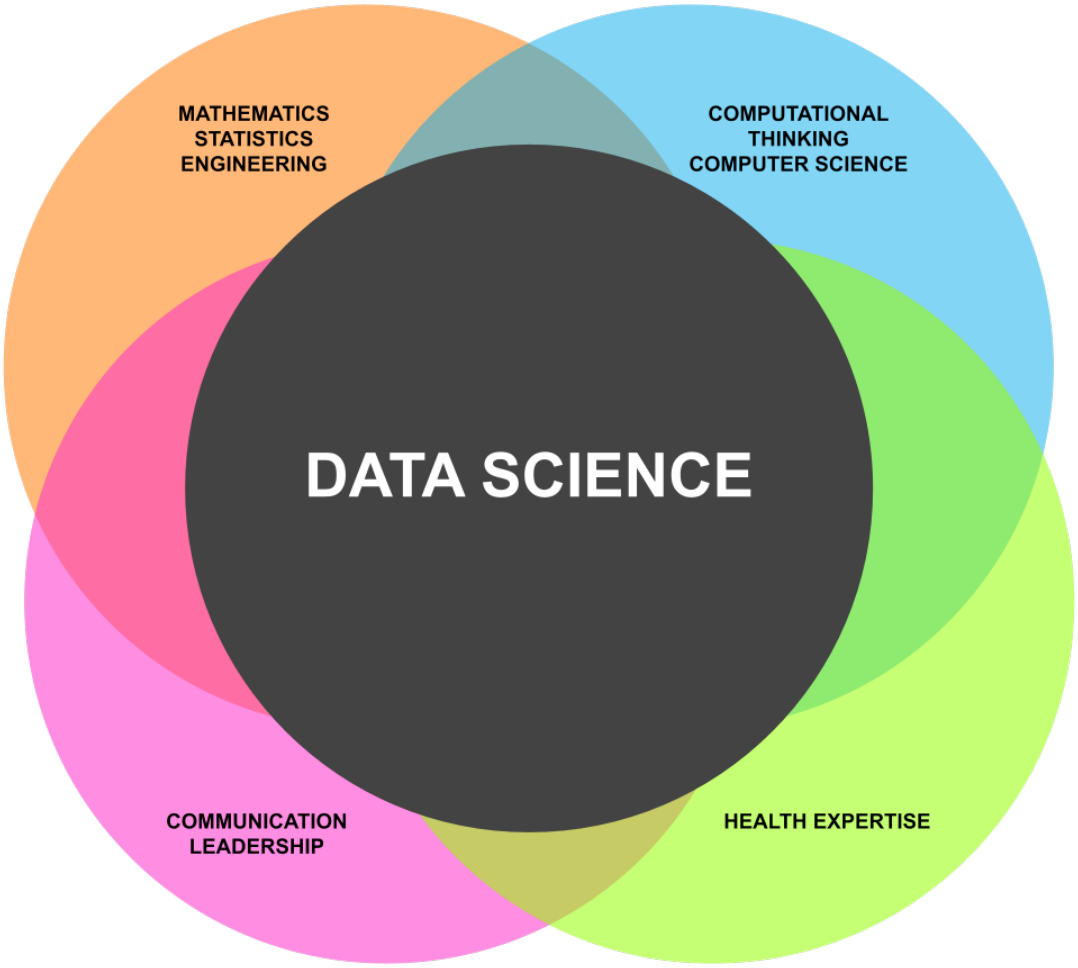
[2]:
```
from plotly import express, io
```

[3]:
```
io.templates.default = "gridon"
```

1.2 Introduction

This chapter serves as a first introduction to health data science. Health Data Science is a multidisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured health data. It combines aspects of data science, statistics, machine learning, and health informatics to generate insights that can be used to improve health outcomes, enhance patient care, and inform health policy.

The image below visualizes the intersection of the major disciplines that form the foundation of health data science. The intersection of these disciplines is where health data science resides.



While data science pertains to all fields of the management and analysis of data, health data science is specifically concerned with the management and analysis of health data. Health data science is a broad field that encompasses many different disciplines, including biostatistics, epidemiology, health informatics, and machine learning.

1.3 The growth in health data science

Health data science has seen significant growth over the past few years. This growth has been driven by several factors.

1. **Increased Data Availability**: The proliferation of electronic health records (EHRs), wearable technology, and other digital health tools has led to an explosion in the amount of avalibale health-related data. Health-related data can be analyzed to gain insights into patient health, disease trends, treatment effectiveness, and more.
2. **Technological Advancements**: Advances in technologies such as machine learning, artifi-

cial intelligence, and cloud computing have made it possible to analyze large, complex health datasets. These technologies can identify patterns and trends in the data that would be difficult, if not impossible, to detect manually or through standard statistical analysis.

3. **Demand for Personalized Medicine**: There iss an ever-growing demand for personalized medicine, which tailors treatment to the individual patient based on their unique genetic makeup, lifestyle, and environment. Health data science plays a crucial role in personalized medicine by analyzing patient data to identify individual risk factors, predict disease progression, and determine the most effective treatments.
4. **Public Health Needs**: The COVID-19 pandemic has underscored the importance of health data science in public health. Health data scientists have played a key role in tracking the spread of the virus, identifying risk factors for severe disease, and evaluating the effectiveness of various interventions.
5. **Policy and Investment**: Governments and private sector companies around the world are investing heavily in health data science. They recognize its potential to improve patient care, reduce healthcare costs, and drive innovation in the healthcare industry.

Given that this is a course on health data science, it serves as a good example to use data to show the current growth in health data science. The PubMed repository is a good source of data for this purpose. PubMed is a free search engine that provides access to over 32 million citations and abstracts from biomedical and life sciences journals. It is maintained by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine (NLM), which is part of the National Institutes of Health (NIH).

The PubMed repository was accessed in mid-2023 and the search term *data science* was used. At the top-left of the first results page is a downloadable link to a spreadsheet file (in comma-separated values file format). This file has two columns: *Year* and *Count*. The *Year* column contains the year and the *Count* column contains the number of results for that year. The file was downloaded and saved as `PubMed_Data_Science.csv` in the same folder as this notebook. The pandas `read_csv` function is used below to import the file as a pandas dataframe object, assigned to the computer variable `df`.

[4]:
```
# Import the PubMed_Data_Science.csv file and assign the dataframe object to␣
↪the variable df
df = read_csv("PubMed_Data_Science.csv")
```

The `head` method is used to display the first five rows of the dataframe.

[5]:
```
# Display the first 5 rows of the dataframe
df.head()
```

[5]:

	Year	Count
0	2022	8837
1	2021	7104
2	2020	4511
3	2019	2686
4	2018	1565

The plotly data visualization package is used to create a bar plot of the number of PubMed results for the search term *data science* over time.

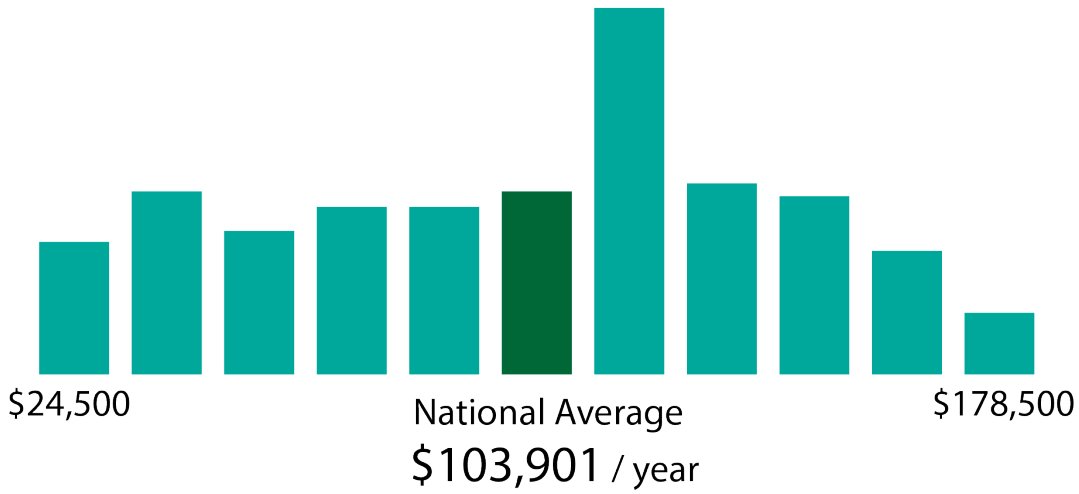
[6]:
```
# Create a plotly bar plot using the Year as the x-axis and the Count of␣
  ↪articles as the y-axis
fig = express.bar(df, x="Year", y="Count", title="Count of Articles using the␣
  ↪search term Data Science in PubMed by Year")
fig.show()
```

From this data, it is clear that, at least in the biomedical and life sciences fields, the number of publications on data science has increased dramatically over the past few years.

1.4 Careers in health data science

The salaries for jobs in health data science is very high. The 2020 Burtch Works Study on Data Science Tools found that the median base salary for data scientists in the United States was $130,000. The median base salary for data scientists in the United States with a graduate degree was $150,000. The median base salary for data scientists in the United States with a graduate degree and 10 or more years of experience was $250,000.

In the image below is a reproduction of a bar plot from ZipRecruiter, showing more specific detail on health data science-related salaries.



1.5 Health data

Health data refers to any data that is related to a person's health status, their health care, or their health determinants. It encompasses a broad range of information types, including but not limited to the following.

1. **Clinical Data**: This includes medical histories, laboratory test results, radiology images, and other data types that are typically found in electronic health records (EHRs). Clinical

data is usually generated during the process of patient care within a healthcare setting such as a hospital, clinic, or doctor's office.

2. **Genomic Data**: This is data derived from an individual's genetic material (DNA). With the advent of technologies like high-throughput sequencing, it's now possible to generate detailed data on an individual's entire genome. Genomic data can provide insights into an individual's risk of developing certain diseases and their likely response to different types of treatment.

3. **Patient-Generated Data**: This is health-related data that is created, recorded, or gathered by individuals, family members, or caregivers to help address a health concern. This can include data from wearable devices (like heart rate or step count), self-reported symptoms, or health diary entries.

4. **Social Determinants of Health Data**: This includes data about the conditions in which people are born, grow, live, work, and age. Factors such as socioeconomic status, education, neighborhood and physical environment, employment, and social support networks, as well as access to health care can influence a wide range of health outcomes.

5. **Claims and Cost Data**: This includes data from health insurance claims, which can provide information on patient diagnoses, procedures, medications, and the cost of care.

6. **Pharmaceutical and Research Data**: This includes data from clinical trials, drug research, and other pharmaceutical research. This data is crucial for the development of new treatments and therapies.

Health data can be used for a variety of purposes, such as improving patient care, conducting medical research, informing public health initiatives, and guiding health policy decisions. However, it's important to handle health data responsibly due to privacy and security concerns.

1.6 Planning a health data science project

Health data science is a rapidly evolving field that leverages the power of data to improve healthcare outcomes. The planning of a health data science project involves several critical steps, including, among other steps and not necessarli in any particular order, defining a research question or questions, developing a research protocol, securing funding, obtaining ethical approval if required, identifying data sources, capturing and wrangling data (that is cleaning up the data, tarnsofrming it into a relevent format, and verifying the integrity of the data), analyzing data, reporting results, and disseminating or communicating the findings.

1.6.1 Research Questions

The first step in planning a health data science project is defining the research question. This question should be specific, measurable, achievable, relevant, and time-bound. This spells the well-know acronymn SMART. A resaerch question should address a gap in the current knowledge or offer a novel approach to a known problem. The research question guides the entire project and influences the choice of data sources, the design of data collection tools, and the selection of data analysis methods.

The topic is the first lecture of the course Research Methods Foundation at the Milken Institute School of Public Health.

1.6.2 Research Protocol

Once the research question is defined, the next step is to develop a research protocol. This document outlines the project's objectives, methodology, and timeline. It includes details about the study design, the data to be collected, the data analysis plan, and the expected outcomes. The research protocol serves as a roadmap for the project, guiding its implementation and helping to ensure that the research is conducted systematically and ethically.

A template for such a protocol is included in this course. In general, the following serves as a guideline for inclusion in a research protocol.

1. **Title**: A clear and concise title that accurately reflects the nature of the study.
2. **Background and Rationale**: This section should provide a brief overview of what is known about the research topic, identify gaps in the current knowledge, and explain why the research is needed.
3. **Objectives**: Clearly defined primary and secondary objectives of the study. These should be specific, measurable, achievable, relevant, and time-bound (SMART).
4. **Study Design**: A description of the study design (e.g., randomized controlled trial, cohort study, case-control study, cross-sectional study, etc.), including the rationale for choosing this design.
5. **Study Population and Sampling**: Detailed information about the study population, including inclusion and exclusion criteria, and the method of participant recruitment and selection.
6. **Data Collection Methods**: A description of how data will be collected, including the type of data (e.g., demographic data, clinical data, survey responses), the data collection instruments (e.g., questionnaires, medical records), and the procedures for data collection.
7. **Data Analysis Plan**: A detailed plan for how the data will be analyzed, including the statistical methods that will be used.
8. **Ethical Considerations**: A discussion of the ethical issues related to the study, including how participant consent will be obtained, how participant confidentiality will be protected, and how potential risks and benefits will be balanced.
9. **Timeline**: An estimated timeline for the different stages of the research project, from participant recruitment to data analysis and reporting.
10. **Budget**: An estimated budget for the research project, including the costs of personnel, data collection, data analysis, and dissemination of results.
11. **Dissemination Plan**: A plan for how the results of the research will be disseminated, including potential journals for publication and conferences for presentation.
12. **References**: A list of references for all sources cited in the protocol.

A well-written research protocol is crucial for the success of a healthcare research project. It provides a roadmap for the project, ensures that the research is conducted systematically and ethically, and facilitates communication about the project with stakeholders, funding bodies, and ethical review boards.

1.6.3 Funding

Securing funding is a crucial step in the planning process. Funding can come from various sources, including government agencies, non-profit organizations, and private companies. The funding proposal should clearly articulate the project's significance, objectives, methodology, and potential impact. It should also include a detailed budget that outlines the project's expected costs.

There are numerous organizations in the United States that provide funding for healthcare research projects. Some of the most prominent organizations are listed below.

1. **National Institutes of Health (NIH)**: The NIH is the largest public funder of biomedical research in the world. It provides funding for research that aims to enhance health, lengthen life, and reduce illness and disability.
2. **Centers for Disease Control and Prevention (CDC)**: The CDC provides funding for a wide range of health research, particularly in areas related to public health and disease prevention.
3. **Patient-Centered Outcomes Research Institute (PCORI)**: PCORI funds research that can help patients and those who care for them make better-informed decisions about the healthcare choices.
4. **Agency for Healthcare Research and Quality (AHRQ)**: AHRQ provides funding for research that aims to improve the quality, safety, efficiency, and effectiveness of healthcare.
5. **Robert Wood Johnson Foundation (RWJF)**: RWJF is the nation's largest philanthropy dedicated solely to health. It provides funding for research and initiatives to help everyone in America have an equal opportunity to live the healthiest life possible.
6. **Bill & Melinda Gates Foundation**: While much of its focus is on global health, the Gates Foundation also funds research in the U.S. that addresses health inequities and improves access to healthcare services.
7. **American Cancer Society (ACS)**: ACS provides funding for a wide range of research projects aimed at understanding and treating cancer.
8. **American Heart Association (AHA)**: AHA funds research related to cardiovascular disease and stroke.
9. **Susan G. Komen Foundation**: This foundation provides funding for research focused on breast cancer.
10. **Pharmaceutical Companies**: Many pharmaceutical companies have grant programs that fund research related to their therapeutic areas of interest.

These are just a few examples. The specific organization that researchers might approach for funding would depend on the nature of their project and the alignment with the funding organization's priorities and interests.

1.6.4 Ethical Review Boards

Before most project can begin, they must be reviewed and approved by an ethical review board. This board ensures that the project complies with ethical guidelines and that the rights and welfare of the participants are protected. The review process involves submitting an application that details the

project's objectives, methodology, potential risks and benefits, and measures to protect participant confidentiality and privacy.

Institutional Review Boards (IRBs), also known as ethical review boards, play a pivotal role in ensuring the ethical conduct of research involving human subjects. Their primary responsibility is to protect the rights, welfare, and well-being of research participants.

IRBs originated in response to historical abuses in human subjects research, such as the infamous Tuskegee Syphilis Study. Today, they serve as an essential checkpoint in the research process, particularly in biomedical and behavioral research.

An IRB is typically composed of a diverse group of individuals, including scientists, non-scientists, and community members. This diversity ensures a comprehensive review of research protocols from various perspectives. Scientists contribute their technical expertise, non-scientists provide a lay perspective, and community members represent the interests and values of the community from which research participants may be drawn.

Before a research study involving human subjects can commence, the study protocol must be reviewed and approved by an IRB. The IRB reviews the protocol to ensure that the study is designed to minimize potential harm to participants, that risks are outweighed by potential benefits, and that participants will be selected in a fair manner. They also ensure that participants will give informed consent, meaning they will be adequately informed about the study's purpose, procedures, risks, benefits, alternatives, and their rights as participants.

Informed consent is a cornerstone of ethical research. It respects individual autonomy by ensuring that participants voluntarily agree to participate in research, fully understanding what participation entails. The IRB reviews the informed consent documents and procedures to ensure they are appropriate and comprehensible to potential participants.

IRBs also conduct ongoing reviews of approved studies to ensure they continue to meet ethical standards. They can require modifications to study protocols, suspend studies that are not being conducted in accordance with the approved protocol, or terminate studies that have been associated with unexpected serious harm to participants.

IRBs play a critical role in upholding ethical standards in research involving human subjects. They protect the rights and welfare of research participants, ensure informed consent, and promote ethically sound research practices, thereby fostering public trust in research.

1.6.5 Sources of Data

Identifying appropriate sources of data is a key step in the planning process. Data can come from various sources, including electronic health records, health insurance claims, patient surveys, wearable devices, and public health databases. The choice of data sources depends on the research question and the available resources.

Open data is data that has been deindentified and made freely available to the public. It can be used for a variety of purposes, including research, education, and innovation. Open data is often used in health data science projects because it is readily available and can be used to answer a wide range of research questions.

There are several sources of open healthcare data available in the United States. The list below provides a brief overview of some of the most prominent sources.

1. **HealthData.gov**: This is the U.S. government's principal health data repository. It provides access to over a thousand datasets on a wide range of health topics, including healthcare quality, health outcomes, medical devices, and more.
2. **Centers for Medicare & Medicaid Services (CMS)**: CMS provides access to a variety of datasets related to Medicare and Medicaid services, including data on utilization, payment, and quality of care.
3. **National Center for Health Statistics (NCHS)**: NCHS is a part of the Centers for Disease Control and Prevention (CDC) and provides statistical information that guides actions and policies to improve the health of the American people.
4. **Behavioral Risk Factor Surveillance System (BRFSS)**: The BRFSS is a system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services.
5. **ClinicalTrials.gov**: This is a database of privately and publicly funded clinical studies conducted around the world. It provides information about the purpose of each trial, who may participate, locations, and phone numbers for more details.
6. **The National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) Program**: The SEER Program provides information on cancer statistics in an effort to reduce the cancer burden among the U.S. population.
7. **The National Health and Nutrition Examination Survey (NHANES)**: NHANES is a program of studies designed to assess the health and nutritional status of adults and children in the United States.
8. **FDA's OpenFDA**: OpenFDA provides APIs and datasets for drug, device, and food-related data.

These sources provide a wealth of information that can be used for a variety of purposes, from academic research to policy development to public health initiatives. However, it's important to note that while these datasets are publicly available, they must still be used responsibly and in accordance with any applicable terms of use.

1.6.6 Tools for Data Capture

Depending on the data sources, different tools may be used for data capture. These can range from electronic data capture systems for collecting data from electronic health records, to survey software for administering patient surveys, to application programming interfaces or APIs for accessing data from wearable devices or public health databases.

Still the two most common tools for the manual capturing of data are spreadsheets and databases. Spreadsheets are a popular choice because they are easy to use and can be used for a wide range of tasks, from data entry to data analysis. Databases are another common choice because they offer more advanced features, such as data validation and data integrity checks.

A spreadsheet is a file made of rows and columns that help sort, organize, and arrange data efficiently. It's a type of software application that enables users to store, manipulate, and analyze data in tabular form.

Each cell within the grid of a spreadsheet can contain a number, text, or a formula. A formula is a

command inserted into a cell that carries out calculations using the data in other cells. This feature makes spreadsheets particularly useful for performing complex calculations and data analysis.

Spreadsheets are widely used in various fields such as business, finance, accounting, and science for tasks like financial analysis, budgeting, project management, data analysis, and record keeping. The most commonly used spreadsheet program is Microsoft Excel, but there are many other programs available, such as Google Sheets and Apple's Numbers.

A database is a structured set of data. It is an organized collection of information that can easily be accessed, managed, and updated. Databases can store data about people, products, orders, or anything else of interest to an individual or an organization.

Databases are typically managed by a Database Management System (DBMS), which provides users with the tools to add, edit, and delete data, generate reports, and perform other operations. There are different types of databases, such as relational databases, object-oriented databases, hierarchical databases, and network databases, each with their own structure and type of DBMS.

Examples of database software, or DBMS, include the following.

1. **Oracle Database**: A popular relational DBMS with a wide range of capabilities for small to large enterprises.
2. **MySQL**: An open-source relational DBMS that is widely used for web databases.
3. **Microsoft SQL Server**: A relational DBMS with a variety of editions for different needs, from small applications to large enterprise solutions.
4. **PostgreSQL**: An open-source object-relational DBMS that supports a wide variety of data types and has strong compliance with the SQL standard.
5. **MongoDB**: A leading NoSQL database, which is document-oriented, meaning it stores data in a semi-structured format known as BSON (a binary representation of JSON-like documents).
6. **SQLite**: A self-contained, serverless, and zero-configuration database engine used in embedded systems and small to medium web and desktop applications.
7. **IBM Db2**: A family of hybrid data management solutions designed to provide a robust set of capabilities to handle data and analytics.

These are just a few examples of the many database software options available. The choice of database software depends on the specific needs and resources of the user or organization.

1.6.7 Data Wrangling

Once the data is captured, it often needs to be cleaned and transformed in a process known as data wrangling. This involves dealing with missing or inconsistent data, removing outliers, and converting data into a format suitable for analysis. Data wrangling is a critical step that can significantly impact the quality of the analysis and the validity of the results. Data wrangling is often the most time-consuming part of a health data science project.

In the era of big data, organizations and institutions across various sectors are inundated with vast amounts of data. However, this data often comes in a raw, unstructured, or semi-structured format that is not immediately suitable for analysis.

Some of the important steps in data wrangling are listed below.

1. **Data Discovery**: The first step in data wrangling involves understanding the data you have. This includes identifying the data types, the structure of the data, and any potential issues that might affect the quality of the data.
2. **Data Structuring**: Raw data often comes in a format that is not suitable for analysis. Structuring involves transforming the data into a format that is easier to work with. This could involve reshaping the data, combining multiple datasets, or converting the data into a different format.
3. **Data Cleaning**: This step involves identifying and correcting errors in the data, such as missing values, inconsistent entries, or outliers. Data cleaning is crucial for ensuring the accuracy of the subsequent analysis.
4. **Data Enriching**: Enrichment involves adding new data or variables to the existing dataset to enhance the analysis. This could involve adding demographic data, calculating new variables, or integrating data from different sources.
5. **Data Validating**: The final step in data wrangling is validating the dataset. This involves checking the data for consistency and accuracy, and ensuring that it meets the requirements of the subsequent analysis.

There are various tools available for data wrangling, including programming languages like Python and R, which have packages specifically designed for data wrangling. These tools can help automate many of the tasks involved in data wrangling, making the process more efficient.

1.6.8 Data Analysis

Another crucuial step is to analyze the data. This involves selecting appropriate statistical or machine learning methods to answer the research question. The choice of methods depends on the nature of the data and the research question. The analysis may involve descriptive statistics, inferential statistics, predictive modeling, or other techniques.

Crucial to any data analysis is exploratory data analysis (EDA). This is a first exploration of the information contained within the data. EDA is indeed a critical step in the data analysis pipeline. It is a philosophy or an approach towards understanding data and involves a variety of techniques to summarize, visualize, and interpret data. The primary goal of EDA is to explore the data to uncover underlying structures, extract important variables, detect anomalies and outliers, test underlying assumptions, and develop simple models.

EDA comprises several steps. Some of these include the following.

1. **Summary Statistics**: EDA involves generating summary statistics for the data. This includes measures of central tendency like mean, median, and mode, measures of dispersion like range, variance, and standard deviation, and measures of shape like skewness and kurtosis.
2. **Visualization**: One of the most powerful tools in EDA is data visualization. Graphical representations of data can reveal patterns, trends, and relationships that are not apparent from summary statistics alone. Common types of visualizations used in EDA include histograms, box plots, scatter plots, and correlation matrices.

3. **Identifying Relationships**: EDA involves exploring the relationships between variables. This can be done using correlation coefficients for numerical variables and cross-tabulations or chi-square tests for categorical variables.

4. **Checking Assumptions**: Many statistical tests and models rely on certain assumptions about the data. EDA can help check whether these assumptions are met. For example, a Q-Q plot can be used to check if data is normally distributed.

There are various tools and software available for conducting EDA. Programming languages like Python and R have extensive libraries for data manipulation, statistical analysis, and data visualization, making them popular choices for EDA. Spreadsheet software like Microsoft Excel and Google Sheets also offer basic tools for EDA.

The various methods of data analysis will be discussed throughout this course.

1.6.9 Data Reporting

Reporting research results is a crucial part of the scientific process. It involves summarizing the findings of a research study, interpreting the results in the context of the research question, and discussing the implications of the findings. The goal of reporting research results is to communicate the findings to others, allowing them to understand, evaluate, and build upon the research.

Effective reporting of research results typically involves several key components, listed below.

1. **Introduction**: This section provides the context for the research, including the research question and the significance of the study.

2. **Methods**: This section describes the methodology used in the study, including the study design, data collection methods, and data analysis techniques. This allows others to evaluate the appropriateness of the methods and to replicate the study.

3. **Results**: This section presents the findings of the study. This typically involves a combination of text, tables, and figures to summarize the data and highlight the key findings.

4. **Discussion**: This section interprets the results in the context of the research question and the existing literature. It discusses the implications of the findings, the limitations of the study, and potential directions for future research.

5. **Conclusion**: This section summarizes the key findings and their implications. It provides a clear answer to the research question and highlights the contribution of the study to the field.

To ensure the quality and transparency of research reporting, several guidelines have been developed. These include the following.

1. The **CONSORT guidelines for randomized trials**. More can be found at Guidelines for reporting outcomes in trial reprots.

2. The **STROBE guidelines for observational studies**. More can be found at the STROBE website.

3. The **PRISMA guidelines for systematic reviews and meta-analyses**. More can be found at the PRISMA website.

These guidelines provide a checklist of items that should be included in the research report to ensure comprehensive and transparent reporting.

1.6.10 Dissemination and Communication of Results

The dissemination and communication of the results can involve publishing the findings in a peer-reviewed journal, presenting the results at a conference or sharing the findings with relevant stakeholders. The goal is to ensure that the knowledge gained from the project is shared widely and can be used to inform decision-making in healthcare.

Effective communication is crucial at this stage. The results should be presented in a way that is accessible to the intended audience, whether they are healthcare professionals, policymakers, patients, or the general public. This may involve using visualizations to illustrate the findings, translating technical terms into plain language, and highlighting the key takeaways.

In conclusion, planning a health data science project involves a series of interconnected steps, each of which plays a crucial role in the project's success. By carefully defining the research question, developing a detailed research protocol, securing funding, obtaining ethical approval, identifying appropriate data sources, capturing and wrangling data, analyzing data, reporting results, and effectively disseminating findings, researchers can leverage the power of data to generate insights that improve healthcare outcomes.

1.7 Conclusion

In conclusion, healthcare data science stands at the intersection of multiple disciplines, leveraging the power of data, statistical methods, and advanced computational tools to generate insights that can improve patient care, enhance health outcomes, and inform policy decisions. The field has seen significant growth in recent years, driven by the proliferation of digital health data, advancements in technology, and a growing recognition of the potential of data-driven approaches in healthcare.

From predicting disease outbreaks to personalizing treatment plans, healthcare data science has the potential to revolutionize the way we understand and approach health and disease. However, it also presents new challenges, particularly in terms of data privacy, security, and ethical use of data. As the field continues to evolve, it will be crucial to address these challenges and ensure that healthcare data science is used responsibly and effectively for the benefit of all.

Understanding the basics of healthcare data science can provide valuable insights into this exciting and rapidly evolving field. As we continue to generate and collect more health-related data than ever before, the role of data science in healthcare will only become more important. The future of healthcare is data-driven, and healthcare data science will be at the forefront of this transformation.