Introduction to Business Analytics

Lecture 2: Data Collection

Iegor Vyshnevskyi

Woosong University

March 11/13, 2023

Agenda

1. The value of data for companies
2. Types of data collection
3. Steps of secondary data collection
4. Defining the research questions
5. Types and sources of secondary data
6. Challenges in secondary data collection
7. Methods of accessing secondary data
8. Real-life examples of data collection
9. In-class assignment
10. Home work

1. The Value of Data for Companies

Netflix

In 2017, valued at over $80 billion, **over 100 times its value** when it went public in 2002.

BUT

- Netflix was not the first company to create television and cinematic content.
- The vast majority of Netflix's content is actually owned by its competitors.
- The lion's share of the material available on Netflix is produced by cable companies across the world.

Why is Netflix different from traditional cable companies that offer shows on their own channels?

- Data driven decisions.
- Netflix tracks every customer action on its website, which allows it to understand its audience and target specific niches with accuracy.
- Netflix's use of data helped it create its hit show "House of Cards", which propelled its viewership and proved the success of its online strategy.

UBER

In 2014, Uber’s valuation was a mammoth 40 billion USD, which by 2015 jumped another 50% to reach 60 billion USD.

BUT

There are other cab companies available in every city.

What makes Uber so special?

- Uber does not own the cab fleet or have drivers. Uber’s key asset is data.
- Uber owns all rights to every bit of data from every passenger, every driver, every ride and every route on its network.

The value of data in Uber's success

- Uber's valuation reached $40 billion in 2014 and $60 billion in 2015, largely due to the value of its data assets.
- Uber's "asset-light" model is made possible by its ownership of data from every passenger, driver, ride, and route on its network.

What conclusion can be made from these examples:

Data is a valuable asset that can enable companies to build **sustainable competitive advantages** and **justify high market valuations**.

By leveraging data to understand customers' behavior and preferences, companies can make better decisions about what products and services to offer, which can lead **to increased growth and profitability**.

2. Types of Data Collection

Two Types of Data Collection

Primary Data Collection

the process of collecting original data, directly from the source.

- Data is designed to meet the specific needs of the research project.

Examples: surveys, interviews, offline quizzes, delphi technique, focus groups and observations.

Secondary Data Collection

the process of gathering information that has already been collected and analyzed by others for purposes other than the research at hand.

- Data has already been gathered for a different purpose other than the present research project.

Examples: sales records, industry reports, interview transcripts, APIs, web scraping.

**The Delphi technique%% is a structured process for collecting and summarizing opinions and feedback from a group of experts or stakeholders on a particular topic or question, with the goal of arriving at a reliable and accurate answer or consensus. It typically involves multiple rounds of feedback and anonymous responses to ensure unbiased and reliable results.

Advantage and Disadvantages of Both Types(1)

Primary Data Collection

Advantages

- Data is specific and relevant to the research question.
- Researchers can control the data collection process.
- Researchers can choose the method of data collection.
- Researchers can establish a relationship with respondents.

Limitations

- Can be time-consuming and costly to collect.
- May suffer from respondent bias.
- Sample size may be too small to make generalizations.
- Researcher must ensure the accuracy of the data collected.

Advantage and Disadvantages of Both Types(2)

Secondary Data Collection

Advantages

- Saves time and money.
- Large data sets available for analysis.
- Data has already been collected, so there is no need to go through the entire research process again.
- Can be used to compare data from different sources.

Limitations

- Data may not be specific to the research question.
- May not be up-to-date or relevant.
- Researchers do not have control over the original data collection process for secondary data collection since the data has already been collected by someone else.
- May suffer from biases introduced by the original data collectors.

In our lecture, we focus mainly on secondary data collection.

3. Steps of Secondary Data Collection

Steps of Secondary Data Collection

1. Define the research question. Determine the specific information needed and the research objectives. This will help to identify the types of data required and the sources that should be explored.

2. Identify potential sources. Search for sources of secondary data that can provide the information needed.

3. Evaluate the sources. Determine if the data is relevant and suitable for the research objectives.

4. Select the sources. Choose the most appropriate sources for the research objectives. Consider the cost, time, and effort required to access and analyze the data.

5. Obtain the data. Collect the data from the selected sources. This may involve downloading data from online sources, purchasing data from a vendor, or obtaining permission to access proprietary data.

6. Organize and store the data. Organize the data into a format that is easy to use and analyze. This may involve cleaning and transforming the data, removing duplicates and errors, and formatting the data for analysis.

7. Document the data. Keep track of the sources of the data, including the dates, locations, and authors. This will help to ensure the accuracy and reliability of the data and will also allow for easy citation of the sources in reports or other publications.

4. Defining the Research Questions

Defining the research questions

to make data-driven decisions

Why is it important to answer right questions?

People say: “To ask the right question, you need to know 80% of the answer”

- Well-defined questions assure the access to obtain relevant and actionable insights from data, and lead to informed and effective decision-making.

Steps to Define the Research Question

1. Identify what problem you want to solve
2. Formulate the effective research question

Identify what problem you want to solve

1. Making predictions

2. Categorizing things

3. Spotting something unusual

4. Identifying themes

5. Discovering connections

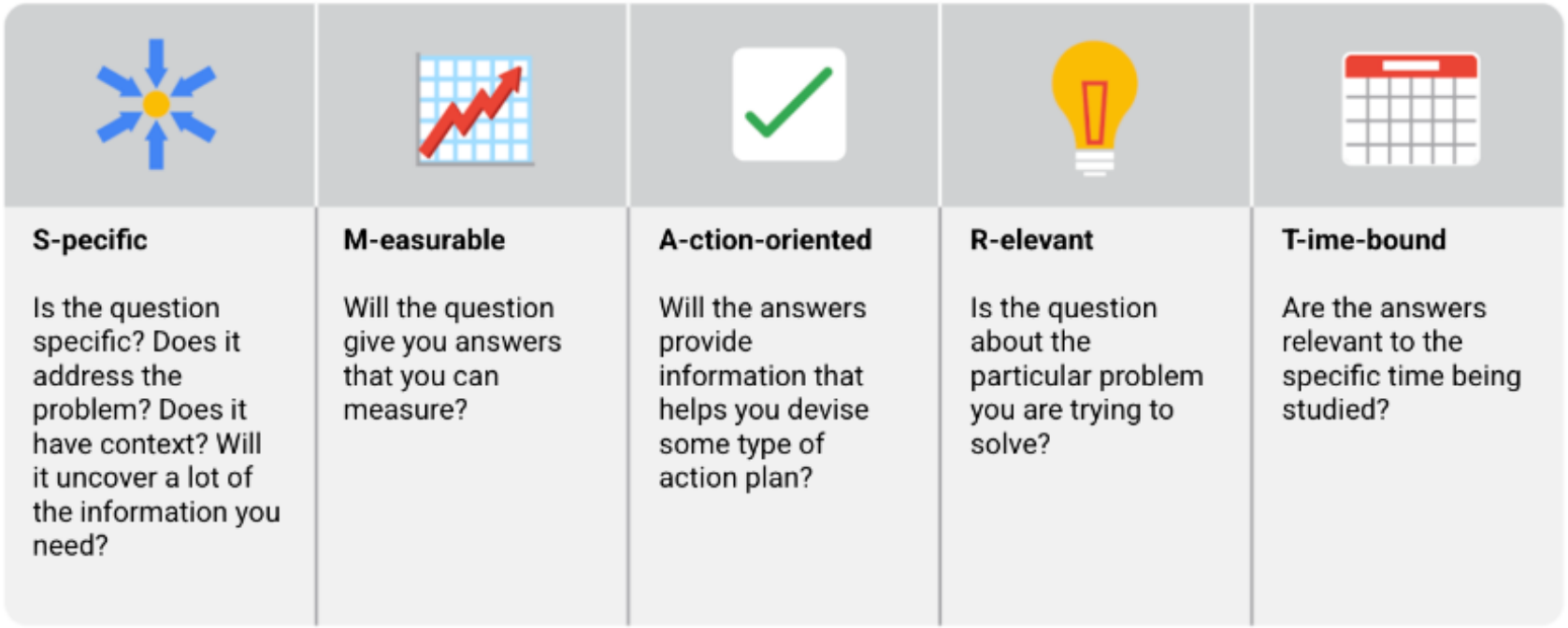
6. Finding patterns

Source: Coursera course “Ask Questions to Make Data-Driven Decisions”

Formulate the effective research question

SMART questions-approach

turns a problem question into one or more SMART questions



S-pecific	M-easurable	A-ction-oriented	R-elevant	T-ime-bound
Is the question specific? Does it address the problem? Does it have context? Will it uncover a lot of the information you need?	Will the question give you answers that you can measure?	Will the answers provide information that helps you devise some type of action plan?	Is the question about the particular problem you are trying to solve?	Are the answers relevant to the specific time being studied?

Source: Coursera course “Ask Questions to Make Data-Driven Decisions”

Example: *How does customer satisfaction affect repeat business in the restaurant industry?*

Specific: Does the question focus on a particular aspect of customer satisfaction, such as food quality or customer service?

Measurable: Does the question include a metric for measuring customer satisfaction, such as customer feedback or ratings?

Action-oriented: Does the question provide insights into how restaurants can improve customer satisfaction and increase repeat business, such as by improving their menu or training their staff?

Relevant: Does the question identify which factors of customer satisfaction are most important in driving repeat business, such as overall experience or value for money?

Time-bound: Does the question consider how customer satisfaction has changed over time and how it has affected repeat business in the restaurant industry in recent years?

5. Types and Sources of Secondary Data

Types of Secondary Data

Internal data: data that is generated and collected within an organization, such as sales records, financial data, and employee records.

External data: data that is collected from sources outside of an organization, such as government reports, industry publications, and market research reports.

Published data: data that is made available to the public through books, journals, and other publications.

Unpublished data: data that is not made publicly available, such as proprietary data, internal company reports, and personal communications.

Sources of Secondary Data

Government sources: census data, surveys, statistical databases, and reports published by government agencies.

Commercial sources: market research reports, industry reports, financial reports, and data from information services.

Non-profit sources: research reports, academic papers, and data from non-profit organizations.

Online sources: websites, social media, databases, and online publications.

6. Challenges in Secondary Data Collection

Major Challenges in Secondary Data Collection

- Data Quality
- Data Relevance
- Data Availability
- Data Limitations

Data Quality

What it refers to: The accuracy, completeness, and reliability of the data.

Why it is important: Poor data quality can lead to incorrect insights and decisions, which can be costly and damaging for organizations.

Examples of pure data quality: Inaccurate data, missing values, inconsistent data, duplicated data, etc.

Methods for evaluating: Checking the source of the data, assessing the methodology used to collect the data, considering potential biases in the data, and conducting statistical tests to identify errors and inconsistencies.

Possible ways to improve the situation: Using data cleaning and standardization techniques, implementing data quality checks during the data collection process, and choosing reputable data sources.

Data Relevance

What it refers to: The degree to which the data is useful for the intended research or analysis.

Why it is important: Collecting and analyzing irrelevant data can waste time, money, and resources, and can lead to incorrect conclusions.

Examples of pure data relevance: Data that is not directly related to the research question, outdated data, or data that is too broad or too narrow for the research needs.

Methods for evaluating: Conducting a thorough review of the research question and objectives to determine what data is needed, and carefully selecting and evaluating the data sources to ensure they meet those needs.

Possible ways to improve the situation: Clearly defining the research question and objectives, conducting a thorough review of available data sources before starting the data collection process, and using targeted and specific data collection techniques.

Data Availability

What it refers to: The ability to access and obtain the desired data.

Why it is important: Limited data availability can hinder research and analysis, and may lead to incomplete or inaccurate conclusions.

Examples of pure data availability: Data that is not publicly available, or data that requires permission or payment to access.

Methods for evaluating: Researching the availability of the desired data before beginning the data collection process, and assessing the cost and effort required to obtain the data.

Possible ways to improve the situation: Using publicly available data sources, negotiating access to restricted data, and considering alternative data sources or methods of data collection.

Data Limitations

What it refers to: The inherent limitations and biases of the data.

Why it is important: Understanding the limitations and biases of the data is important for interpreting and analyzing the data accurately.

Examples of pure data limitation: Biases in the data due to the sampling method used, limitations in the scope of the data, or limitations in the time period covered by the data.

Methods for evaluating: Understanding the limitations of the data source, and assessing the potential biases of the data through statistical analysis or other methods.

Possible ways to improve the situation: Using multiple data sources to complement each other and fill gaps in the data, and carefully documenting any limitations or biases in the data.

7. Methods of Accessing Secondary Data

Online databases: accessing data through online resources, such as databases and data repositories. Examples: FRED (Federal Reserve Economic Data),Bloomberg Terminal, PitchBook, S&P Global Market Intelligence, Statista, IBISWorld

Web scraping: using software to extract data from websites. Examples: BeautifulSoup (Python library), Scrapy (Python framework), RSelenium (R interface to Selenium), rvest (R package), Octoparse (web-based tool), WebHarvy (Windows desktop app), ParseHub (web-based tool)

API: accessing data through Application Programming Interfaces provided by data sources. Examples: Twitter API, Facebook API, and Google Maps API.

Web search: searching for data on the internet using search engines. Examples: Google, Bing.

Social media monitoring: collecting data from social media platforms. Examples: Twitter, Facebook, and Instagram.

Electronic surveys: conducting surveys online through email, websites, or social media. Examples: SurveyMonkey, Google Forms, and Qualtrics.

Electronic health records: accessing and analyzing medical data electronically, such as patient health information and medical histories.

8. Real-Life Examples of Data Collection

Basic web-scraping with R

1. Install and load rvest package.

- Rvest package is one of the most popular R packages for web-scraping.

Code:

```
install.packages("rvest")

library(rvest)
```

2. Set the `url` variable to the URL of IMDb's top-rated movies list.

Code:

```
url <- "https://www.imdb.com/chart/top"
```

3. Sets the node that will select the movie titles on the IMDb webpage.

- Choosing the correct elements on a webpage can be done using the web inspector tool in your browser. In Google Chrome, for example, you can right-click on the element you are interested in and select "Inspect" to open the web inspector. This will show you the HTML code of the page, and you can then hover over different parts of the code to see which element is being highlighted on the page.
- Once you have identified the element you want to scrape, you can right-click on the element in the web inspector and select "Copy selector" or "Copy XPath" to copy the CSS selector or XPath expression for that element. You can then use this selector or expression in your web scraping code to extract the desired information from the page.
- It's important to note that the structure of a webpage's HTML code can change over time, so you may need to update your selector or expression if the page's structure changes.

Code:

```
titles <- read_html(url) %>% html_nodes("#main > div > span > div > div > div.lister > table > tbody > tr:nth-child(n) > td.titleColumn > a") %>% html_text()
```

4. View movies titles

Code:

```
view(titles)
```

Result:

	x
1	쇼생크 탈출
2	Daeboo
3	다크 나이트
4	The Godfather Part II
5	12명의 성난 사람들

5. Web-scrap year and rating info

Code:

```
years <- read_html(url) %>% html_nodes("#main > div > span > div > div > div.lister > table > tbody > tr:nth-child(n) > td.titleColumn > span") %>% html_text()

ratings <- read_html(url) %>% html_nodes("#main > div > span > div > div > div.lister > table > tbody > tr:nth-child(n) > td.ratingColumn.imdbRating > strong") %>% html_text()
```

6. Combine the extracted data into a data frame and view the results

```
imdb_data <- data.frame(title = titles, year = years, rating = ratings)

view(imdb_data)
```

Result:

	title	year	rating
1	쇼생크 탈출	(1994)	9.2
2	Daeboo	(1972)	9.2
3	다크 나이트	(2008)	9.0
4	The Godfather Part II	(1974)	9.0
5	12명의 성난 사람들	(1957)	9.0

Web-scraping of the text

1. Set the `url` variable to the respective URL (Woosong university web-site/About/Who we Are).

Code:

```
url <- "https://english.wsu.ac.kr/page/index.jsp?code=eng0101"
```

2. Set the node and view.

Code:

```
who_we_are <- read_html(url) %>% html_nodes("#rightCont > div") %>% html_text()

print(who_we_are)
```

Result:

```
[1] "\r\n\tWoosong University belongs to the Woosong Education Foundation which was esta
blished in 1954 and has a long history of outstanding work in education, teaching, and t
raining. Founded by the late Kim Jung-Woo, he established a vision for students and decl
ared that\r\n\r\n\t\r\n\t\t“We call on our students to devote themselves to society, acc
epting the continual search for the truth. Encouragement of personal growth assures that
 our graduates will have the skills and the motivation they can be proud of. Our school
 will always uphold the same ideals we foster in our students and set new standards for
 education in Korea.”\r\n\t\r\n\r\n\tThis vision of Kim Jung-Woo has become a vision not
 only for Korean students but for all young people in our international community. As we
 face new challenges on a global scale and as technology shrinks our world into one univ
ersal society, we will look forward together.\r\n\r\n\tFounder Kim Jung-Woo’s clear visi
on and passion for education and the tireless efforts of members of the Woosong Educatio
```

9. In-class assignment

10. Home work

Download Postgre SQL from www.postgresql.org and made all proper actions for installation.

Than, download www.dbeaver.io and made all proper actions for installation.