Implementation and Automation of Data Science Workflow and Visualization

William Su

CSBG Summer Intern

06/01/2018 - 07/31/2018

Data Science Intern Final Report

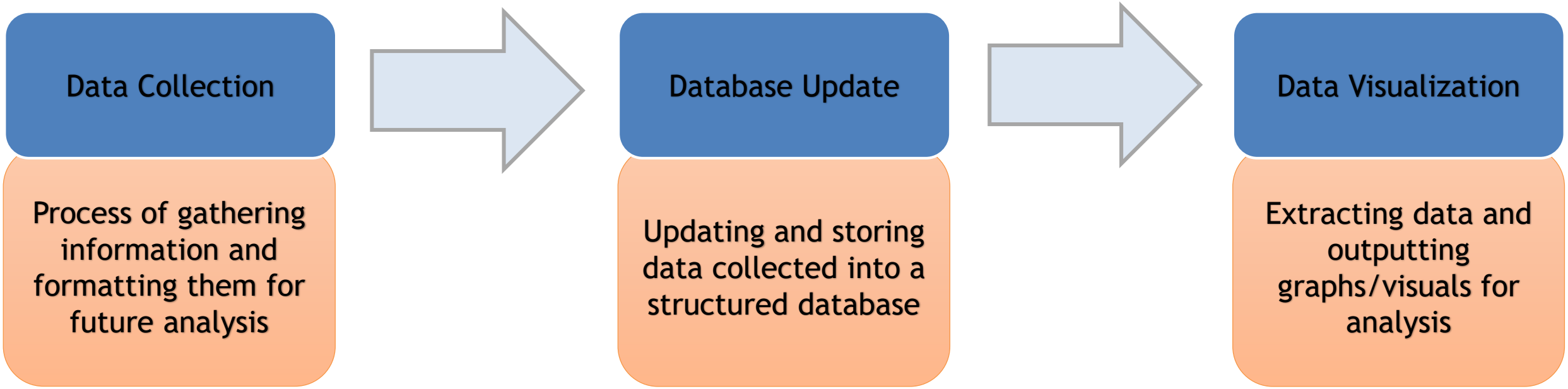
This is a full report of the project I completed and the lessons I learned during my time as a Data Science intern at Lam Research under Mr. DC Lin and Mr. Li Peng's supervision. Below is a table of contents of the upcoming slides.

Data Science Intern Final Report

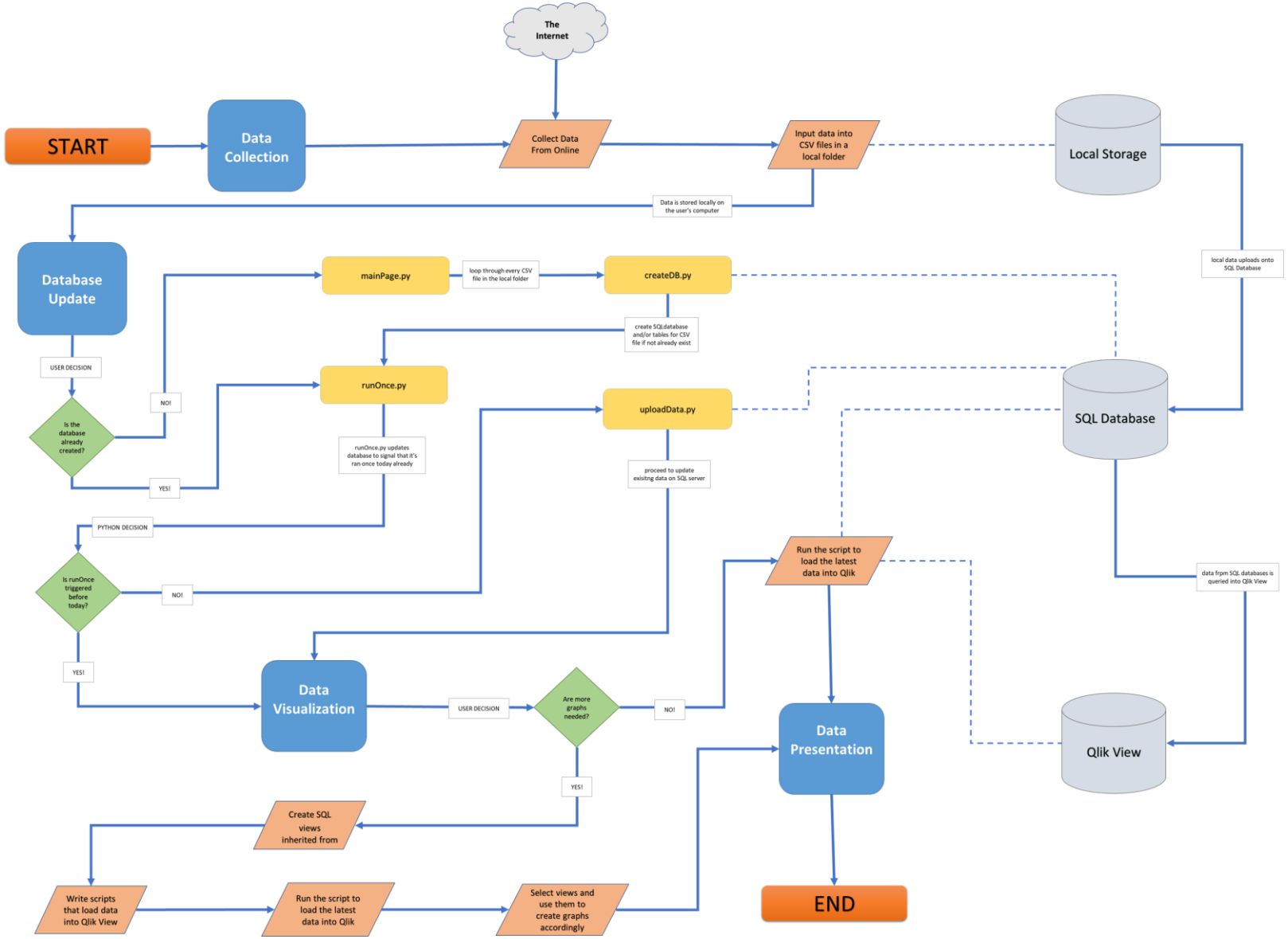
▶ Overview of Data Science Workflow

Central Question:

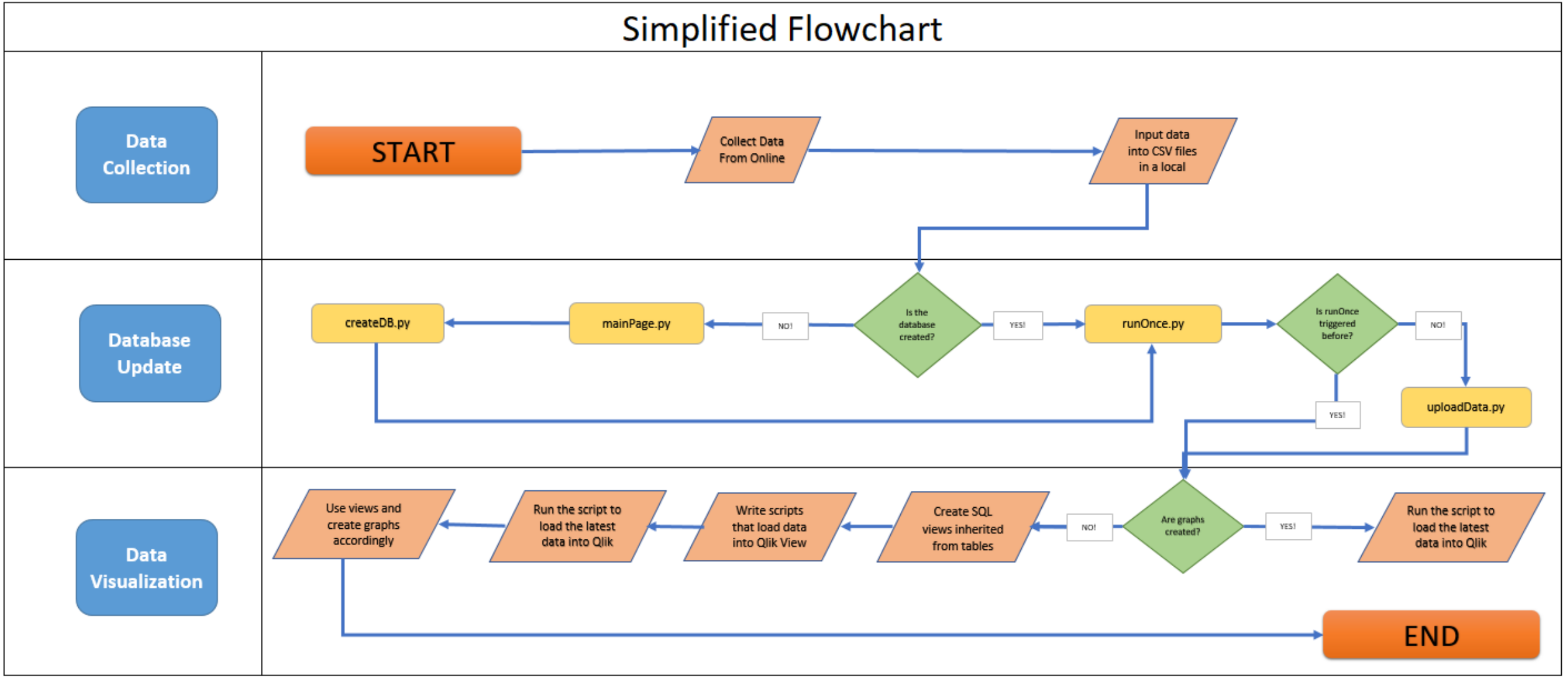
- **How can each step be automated?**



Workflow Process Flowchart

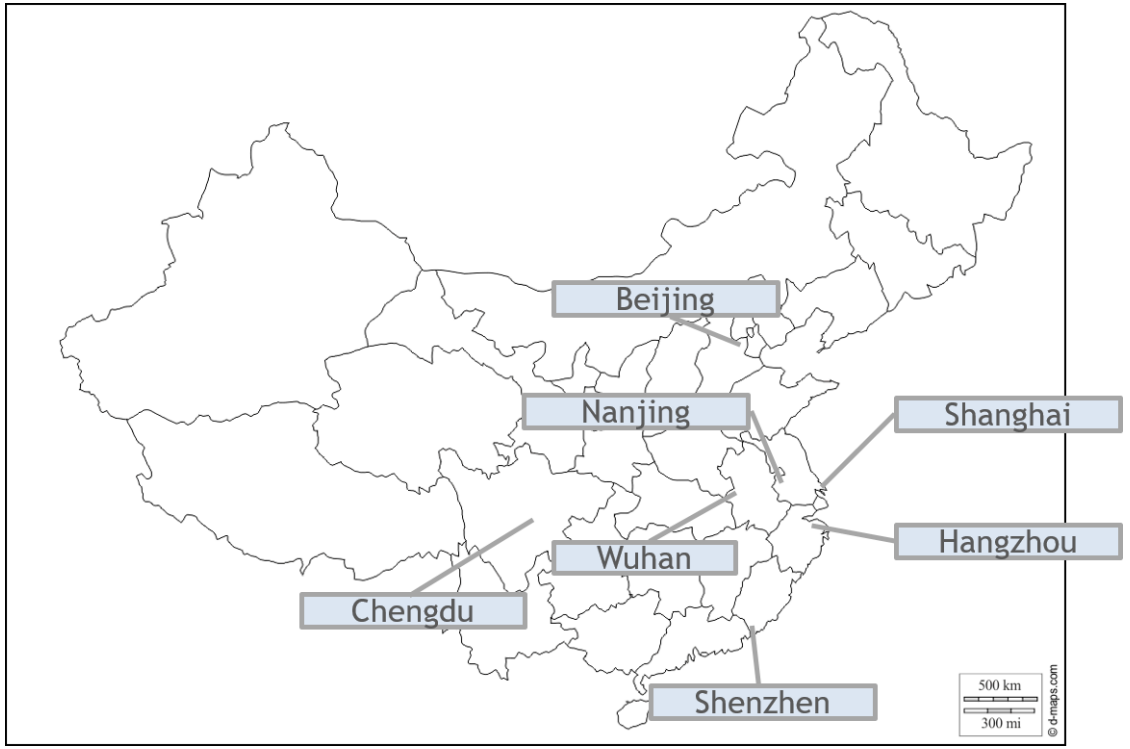


Workflow Process Flowchart



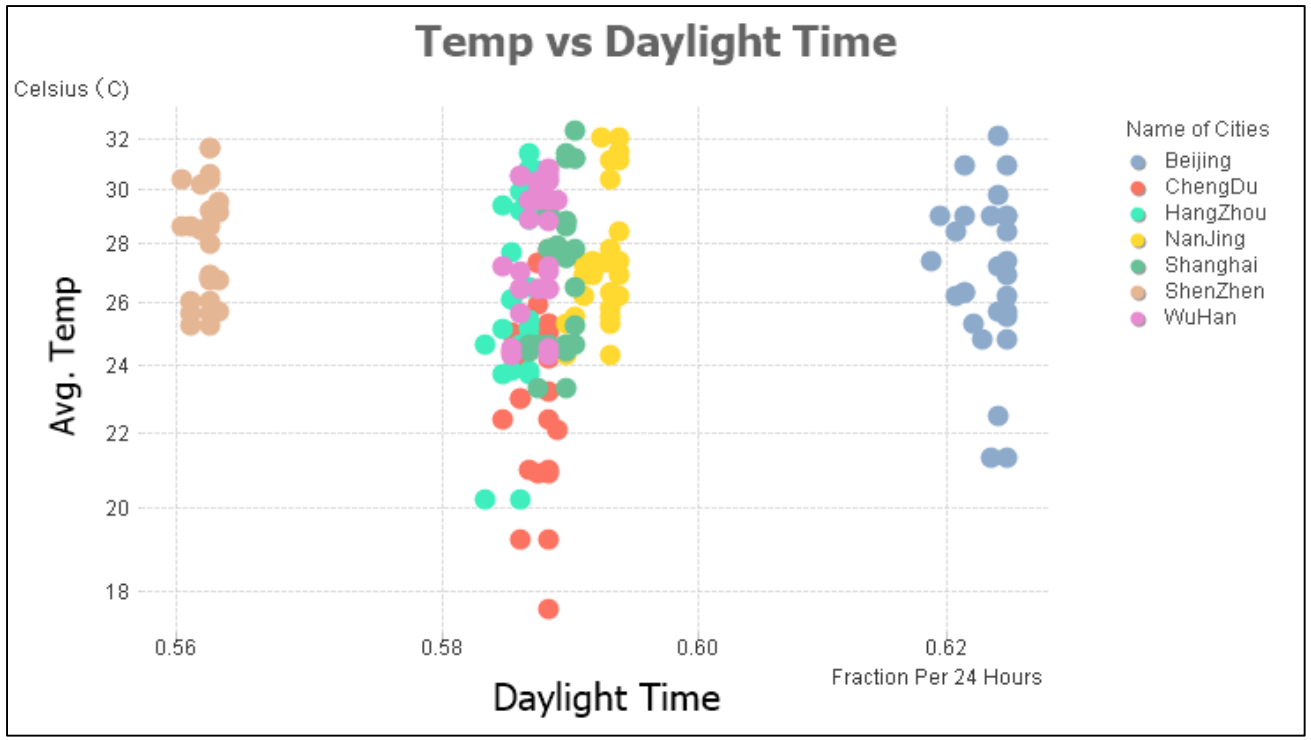
Case Study: Climate in China's Major Cities

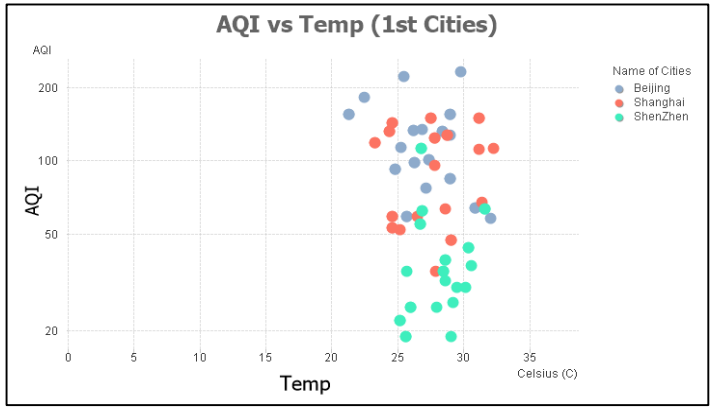
- **Topic: The Climate in 7 China Cities**
  - To practice the entire workflow, I was given a project that has to deal with the daily weather in 7 Chinese cities.
  - Everyday (except a few days), I collected the weather data for each city on my own and inserted them into CSV files.
  - The end goal is to go through the process of data collection, data upload, and data visualization so that ultimately, I can produce graphs that make sense of the climate in every one of these cities.
  - Below are the 7 cities I researched on:
    - Shanghai
    - Beijing
    - Shenzhen
    - Nanjing
    - Wuhan
    - Hangzhou
    - Chengdu

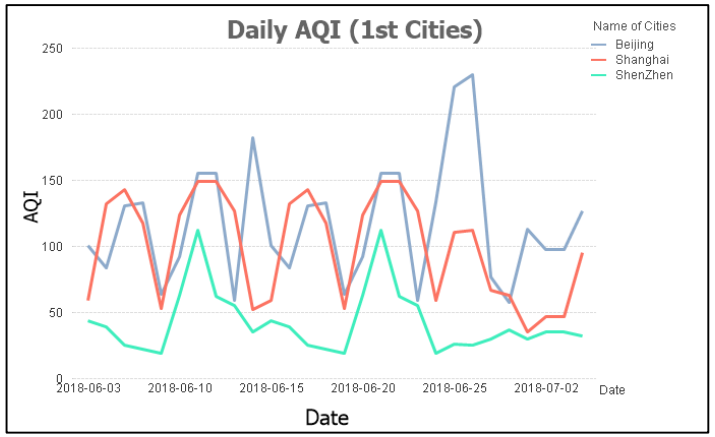


Case Study: Climate in China's Major Cities

- Examples of graphs outputted in Qlik View.
- Data is extracted from SQL database.
- Click to enlarge/minimize the graphs on the right.







Lessons Learned Regarding Data

- **Data Collection**
  1. Data can be inputted in many numerable ways → can lead to major problems later on (example: data fail to upload to database).
  2. Always provide an easy/standardized interface for users.

- **Database Update**
  1. Importance of recording the history that data is added (“historical snapshot”).
  2. Advantages of using database over excel:
     1. Tables can interact with each other.
     2. Can store larger tables.
     3. Cloud/supports multiple user at the same time*.

     *many Excel sheets can support cloud services too today

- **Data Visualization**
  1. Raw data gets confusing. Use tools such as views to organize data for specific graphs.
  2. See through behind the numbers: what’s the story/context behind the data?
     1. Helps with identifying missing/wrong data.
     2. Creates better graphics.
     3. Makes you a better presenter as you know what points to focus on.

Lessons Learned Regarding Data

- **There is no right or wrong in how these three steps can be operated.**
  - Choose the way that best fits what is needed.
    - Current size/complexity of data
    - Potential growth of the size of the data
    - Financial costs

Lessons Learned Regarding Industry

- **PowerPoint Presentations**
  - Make sure the PowerPoints' format is standardized.
  - Highlight key points in all graphics.
- **Paperwork/Formality/Goals**
  - The bigger the company, usually the more paperwork/surveys an employee needs to fill out. This is also because usually the bigger the company, the more guidelines there are.
  - Companies may often ask employees to set yearly, career, or personal goals on some sort of company website. Using these goals as a metric to measure an employee's work performance is very common, especially in bigger companies where there are too many people to keep track of.