
ASEAN
AUSTRALIA
SMART CITIES
TRUST FUND
Asian Development Bank
Australian Government
Department of Foreign Affairs and Trade
ADB
SMART CITIES
Guided Learning Programme
Home Exercise 3
Begin
Home
Exercise
ASEAN Australia Smart Cities Trust Fund (AASCTF)
December 2022

Guided Learning Programme

ASEAN AUSTRALIA SMART CITIES TRUST FUND
Asian Development Bank

BEFORE YOU CONTINUE...

Please download this file onto your computer or mobile device before working on this exercise to avoid overwriting the original file.

Instructions on how to submit your finished work will be provided at the end of this document.

If you encounter any issues with this exercise, please contact your country's AASCTF national focal point.

Next

Guided Learning Programme

ASEAN AUSTRALIA SMART CITIES TRUST FUND
Asian Development Bank

Welcome

Guided Learning Programme

The key to unlocking a smart, livable city is data management.

Most cities today generate ample amounts of data from services and operations that power the function of a city; however, this data is not maximized enough to inform urban planning and decision-making.

The growing concern about data privacy and security also hampers the use of digital systems and data collection.

What can cities do to ensure these vital information are not wasted? How can data governance be improved, and how can it lead to more livable cities?

These are some of the questions that will be addressed in the Guided Learning Programme (GLP) of the ASEAN Australia Smart Cities Trust Fund (AASCTF)

Next

Guided Learning Programme

ASEAN AUSTRALIA SMART CITIES TRUST FUND
Asian Development Bank

The programme structure

This programme is aimed at professionals who are starting their journey working with **urban data** and will be responsible for planning data projects in their city.

The programme consists of three webinars with supplementary coursework to be completed by the specific deadline. You are now to begin the third home exercise.

Home Exercise 1

Data Collection and Storage

Home Exercise 2

Data Governance and Security

Home Exercise 3

Data Analysis and Decision Making

< Back Next >

Home Exercise 3

ASEAN AUSTRALIA SMART CITIES TRUST FUND
Asian Development Bank

Home Exercise 3 – Data Analysis and Decision Making

The third home exercise in the series is called 'Data Analysis and Decision Making'. In this exercise, we will focus on how we combine data with analysis to make improved decision making.

We kindly ask you to reflect on the questions being asked in the home exercise. Note that the output of the course will be for you to have developed a basic urban data project proposal.

< Back Next >

Home Exercise 2

ASEAN AUSTRALIA SMART CITIES TRUST FUND
Asian Development Bank

Types of slides

This home exercise is composed of three types of content: Warm-ups, activities, and reflections.

They are labelled as such on the left sidebar of each slide.

< Back Next >

Guided Learning Programme

ASEAN AUSTRALIA SMART CITIES TRUST FUND
Asian Development Bank

Welcome

Home Exercise 3

In this home exercise, you will be introduced to how data is used for decision making in smart cities, as well as actual methods of data processing and data analysis

- Understand the potential pitfalls e.g. bias, limited data and more.
- Be introduced to actual methods of data processing and data analysis including the data value chain
- Understand how to use data for decision making in smart cities

Type: Training
Module length: 60 min

Start Home Exercise

Understanding the data value-chain: From collection to decision-making

ASEAN AUSTRALIA SMART CITIES TRUST FUND
Asian Development Bank

Introduction to the data value-chain

Ultimately, the goal of collecting data is to support relevant stakeholders in decision-making to achieve the desired vision and goals set for a data project in a smart city. The process from collection to decision-making is described as the data value-chain and can be described by the following overall steps:

1. Data Collection & Data Quality
2. Data Analytic & Information Creation
3. Reporting, Decision-making & Application

It's good practice to outline the aim, scope and methods for the three steps prior to starting data collection. This is, among others, to ensure that the selected analytic tools as well as the quality and scope of the data is representative of the purpose and application of decision-making. Furthermore, it is important to include considerations related to resources and time in realising these steps in the planning of your data project. This is because, the most time-consuming part of the above process is related to the collection and processing of data (step 1).

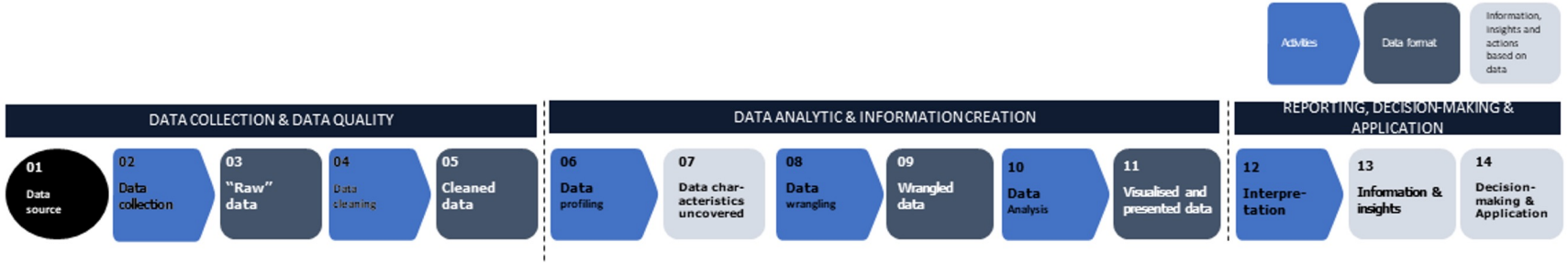
In the previous webinars, the aim, scope and approach for data collection was introduced. Consequently, the following pages will focus on the remaining steps in the data value-chain.

Understanding the data value-chain: From collection to decision-making

ASEAN AUSTRALIA SMART CITIES TRUST FUND
Asian Development Bank

In the previous webinars of the Guided Learning Programme, the aim, scope and approach for data collection was introduced. Consequently, the following pages will focus on the remaining steps in the data value-chain.

The data-value chain specifies the *activities* (light blue box) performed to *process* the data into a desired *format* (dark-blue box) for a given *application* (light-grey box). For example, in the first step "Data Collection & Data Quality", data collection and cleaning is performed to transform the "raw" data collected from the source into "cleaned" data ready for data analytic and information creation.



< Back Next >

Understanding the data value-chain: From collection to decision-making

ASEAN AUSTRALIA SMART CITIES TRUST FUND
Asian Development Bank

What is Data Quality?

Data quality is about ensuring data has sufficient quality for the intended application and use – so-called "fitness of purpose". Different literature provide different characteristics that constitute "data quality". But overall they share similar characteristics based on these two following high-level data quality characteristics:

Validity: Validity is about "measuring what is supposed to be measured". Overall, this includes aspects related to *accuracy*. Accuracy is about the correctness of a data value. For example, temperature sensor can have an accuracy level of about 0.5°C. This means, if the temperature is measured at 20°C, it could in reality be between 19.5-20.5°C. Selecting ICT with a high accuracy is important if an observed effect is expected to be small.

Reliability: Reliability is about *consistency* and minimisation of *bias*. It is important that data is collected and processed in a consistent manner, so the application of the same type of data for similar settings yields similar results. For example, a temperature sensor placed in one room and then moved to another room, which has similar condition as the former, should show the same temperature value. Bias can be introduced by the applied data collection and processing approach. There are different types of biases (non-response bias, systematic offset, respondent bias, selection bias, confirmation bias, etc.). Therefore, it's important to plan the data collection and processing to minimise risk of bias. Moreover, it requires domain experts to be able to assess whether a certain dataset might have biases. Therefore, it is important to be aware of them prior to analysis and interpretation of the results. For example, a project intends to compare a specific performance metric of one group with another, and for this a specific ICT solution is implemented to collect this data. If this ICT solution is used more by the one group over the other, this creates an imbalance in the dataset leading to skewed results and wrong interpretation – thus leading to sampling bias.

Data Quality Assurance

As formerly explained, the level of quality in a dataset depends on its application and purpose. For example, data used for decision-making about citizens lives should have high accuracy, be bias-free and be collected in a consistent manner.

If high data quality is required, ensure to plan your data collection to obtain high quality data. This is a *preventive* type of control in ensuring data quality. However, this can be difficult in terms of time, money and resources when data is collected "in the wild". Additionally, in some cases, you might obtain datasets from other projects or sources. It's therefore important to be aware of any limitations in the dataset. This can be investigated through *data profiling* (see below). This is a *detective* type of control. *Corrective* control is applied for improving the data quality of the collected data. Below an overall framework is described for improving data quality. These can be deployed automatically, manually or as a combination of both.

Data cleaning: Applied as close to the data source as possible. Ensures invalid/corrupted/duplicate data is handled appropriately, and that contextual information of the data is provided.

- Check and assign meta-data (location tags, data source tags, name, hardware ID, etc.) according to defined standards.
- Identify and handle missing data, erroneous/invalid values, duplicates, etc.
- Align timesteps (time-series data).

Data profiling (explained later): Discover through exploratory analysis or comparison to ground-truth data, if available, features and characteristics of the data. For example, the accuracy of data values, or observation of patterns caused by expected biases.

Understanding the data value-chain: From collection to decision-making

ASEAN AUSTRALIA SMART CITIES TRUST FUND
Asian Development Bank

Examples of Data Quality Issues and Solutions

There are different types of data quality issues. In some contexts, some issues can even be relevant information. Consequently, the following provides an overview of some typical issues presented along with some possible solutions:

Invalid data value: Caused by signal problems in the hardware, problems in database set-up or wrong manual data input. These issues can be fixed by applying automatized checks as close to the data source as possible. E.g., proposing data controls for how a name or address is provided.

Lack of data integration: Arises if combining different data models with inconsistent meta-data tags, different timesteps, etc. Some of these issues (e.g., aligning timesteps) can be fixed through data processing. Fixing data integration issues between two or more data models automatically might be difficult. A proper Data Governance Program is therefore important to avoid or limit these type of issues.

Lack of contextual information: If no relevant contextual information is provided, interpretation of the results can become limited. Meta-data tagging becomes important. Meta-data can be descriptive (e.g., location, year), technical (e.g., data model origin, hardware ID) and administrative (e.g., ownership, data creation date).

Duplicates and "expired" data: Duplicates can arise if merging different data sources. This can be avoided by identifying the most up-to-date and correct source, defined as the authoritative data source. It's important to also set-up automatic checks for identifying expired data. E.g., through yearly checks, or by using automatic queries to the authoritative data source, as to ensure the database is populated with the latest data value.

Outliers and scope: These are not necessarily considered as issues in some context. However, in other applications they might skew results and interpretation. Through data profiling these "issues" can be identified. But they should not be removed at the point of origin.

Note that ICT and data platforms might already apply data processing techniques before making the data available to the user. For example, to limit data traffic, some cloud platform providers aggregate the data values to larger timesteps, and it's not always possible to know what type of approach is used. Consequently, this can obliterate some relevant patterns and outliers.

INTRODUCTION

WARMUP

TIME: 5 minutes

REQUIRED MATERIALS: None

Objective of your data project

Please revisit your submission for Home Exercise – And provide here the background, vision and objective of your urban data project. Specify the objective of your data analysis

Example:

Transportation is heavily realigned on fossil fuels and account for 37% of global CO2 emission. Furthermore, it leads to increased pollution causing poor outdoor air quality associated to the cause of poor health and premature death.

The vision of this smart city project is to improve the air quality of the city for the benefit of citizens' health and well-being. The objective of the smart city project is to deploy smart technologies to optimise traffic flow to reduce pollution and CO2 emission from car transportation. The results will help decision-makers and urban planners to improve road and transportation in the city.

The objective of the data analysis is to:

1) *Describe features of "traffic"*
2) *Create a prediction model of the traffic, air pollution and CO2 emission*
3) *Use the model to reduce traffic jams and pollution by implementing different city interventions affecting conditions leading to increased traffic and pollution.*

Understanding the data value-chain: From collection to decision-making

ASEAN AUSTRALIA SMART CITIES TRUST FUND
Asian Development Bank

What is Data Analytic?

Data Analytic is the use of mathematical and statistical tools for obtaining information and insights into a given problem based on data. Data analytic is already conducted during data quality assurance as part of the *data profiling* step. Data analytic can be grouped into three types depending on the application of the data:

Descriptive: Typically applied to investigate obvious patterns in the data. For example, the number of visitors at a given place. The temperature measurement over a given time period. Typical tools applied are simple statistics and visualisation tools such as histograms, box-plots, calculation of means and variance or the use of summary tables. Descriptive analytic is useful as a first step to get familiar with the collected data.

Diagnostic: Applied to investigate the cause and effect of a specific pattern or observation. For example, if there is a correlation between number of visitors and the type of events at a given place. Typical tools are more advanced statistical tools and visualisations such as regression models, correlation tests, scatter plots, hypothesis testing, pattern recognition, etc. Diagnostic analytic is useful to get further information and evidence into expected patterns. It is useful to identify features in the data, develop predictive models or to identify biases or limitations.

Predictive: Applied to predict a specific outcome based on input variables. For example, this can be to predict the future number of visitors in a given place based on the type of events planned to occur. Predictive analytic uses modelling techniques (model creation, testing and validation) such as regression models, non-parametric models or machine learning.

Prescriptive: Using predictive models, prescriptive analytic is applied to determine the best possible outcome based on a given set of criteria and choices. For example, suggesting a favourable day in a week for a specific event to maximise number of people at the event. The applied solutions are artificial intelligence, evolutionary algorithms, stochastic optimisation techniques, etc.

Understanding the data value-chain: From collection to decision-making

ASEAN AUSTRALIA SMART CITIES TRUST FUND
Asian Development Bank

Performing Data Analytic

Data Profiling and Data Wrangling is intended to identify and highlight features in the data for scoping the data analysis and identifying limitations in the data validity and reliability.

Data Wrangling is the process to transform the data to highlight its relevant features. Below are some examples of Data Wrangling techniques:

- Filtering: E.g., removing outliers, focusing on specific time-period or reducing complexity or quantity of data.
- Grouping and Summarising: E.g., Binning of continuous data, grouping similar data patterns
- Calculating: E.g., normalisation, scaling or calculation of new type of data based on the collected data.
- Merging/joining: Combining different datasets.

When you have identified relevant features as well as established the limitation and scope of your data, the next step is to perform the actual data analysis to obtain information and insights related to your objectives and purpose of your smart city project. It is important to formulate some specific questions that can provide evidence and insights into the projects objective or main question/hypothesis.

From collection to decision-making

ASEAN AUSTRALIA SMART CITIES TRUST FUND
Asian Development Bank

Data to decision making

In the context of the smart city and urban projects, decisions are made on three main levels: strategic, tactical and operational. This hierarchical framework can help determine the stakeholders, the type of decisions they are making and .

The strategic level of decision-making is the basis for tactical and operational decisions. The strategic level includes processes and activities for setting long-term goals, policy development, visions, or values for the overall development of the city as a whole.

The tactical level considers medium-term mid-level decisions to achieve the results specified at the strategic level. This level refers to the development of concrete agendas and all actors who regularly deal with programs, funding, and establishment of networks and partnerships. These actors include planners, universities, etc.

Finally, the operational level corresponds to experiments and actions with a short-term vision, which involve the implementation of goals and execution of concrete projects. Operational decisions are mostly used to give operational solutions or to assess results obtained by low-level managers in smart cities.

EXAMPLE
ACTIVITY

TIME: 15 minutes

REQUIRED MATERIALS: None

Understand the Data Value Chain

Provide a brief description of your steps in relation to the data value-chain

Type your answer here…

Who are the decision-makers and what information do they need to make decisions?

Data collection:

Data cleaning:

Data profiling & wrangling:

Data analysis:

Example:
The data analysis is intended to provide urban planners to make tactical decisions for improving roads and transportation in the city. For this, they need to know how the current traffic flow is in the city, how it impacts air quality and what conditions affect the traffic flow.

Data collection: *The deployed smart city technologies will collect cellular data from citizens smartphone, road cameras and measurement data on the air quality in the city. We want to have a large spatial and temporal data coverage in the city for achieving high reliability. Consequently, we use low-cost sensors for obtaining a high data coverage as they are less costly than scientific instruments. Thus, we accept that the deployed sensors have lower accuracy.*

Data cleaning: *Location data is provided as GPS coordinates. Missing data will be removed. Timesteps between different sensors and datasets should be synchronised. Air quality sensors and data from cameras should be assigned GPS coordinates in the similar format as location data.*

Data profiling & wrangling: *We use descriptive and diagnostic analysis to identify relevant features in the data to describe and determine the risk of traffic (e.g., number of cars and time spent in a location). Using diagnostic analysis, we want to investigate the variables affecting the risk of traffic (e.g., time of day, day of the week, weather condition).*

Data analysis: *A predictive model will be developed to predict the air pollution and the risk of traffic based on the identified features. The model will be tested and validated to determine its accuracy and reliability. The model will be used to investigate different scenarios for reducing traffic and air pollution.*

Understanding the data value-chain: From collection to decision-making

ASEAN AUSTRALIA SMART CITIES TRUST FUND
Asian Development Bank

Presenting the results from your Data Analytic Study

There are different methods for presenting your results. Select your type of reporting based on your expected audience, type of data you are presenting and what "story" you want to tell with the data. Below are some examples of different reporting formats:

Graphs & report: This is useful for a more comprehensive dissemination of project outcome and results. Graphs are used to present complex results in a more comprehensive manner without use of extensive texts. Graphs should be simple in terms of presenting patterns and relationships in the data, and contain as much information for understanding the graph (e.g., use of titles, units, legends, colouring). Texts are used to guide the readers in interpreting the graph and results. Useful for technical audience. Graphs are useful to show distribution, relationship between variables, outliers and trends, similarities and differences in data, etc.

Tables: This is useful for technical audience that wants to use relevant results from the study to conduct further analysis. Ensure to provide enough information about the content of the table (e.g., scope, limitation, methods applied).

Key-figures & infographic: This is a useful approach to disseminate to non-technical audience, in a short and summarised manner using graphics and key figures. It is important to provide some few notes in terms of scope and limitation as well as a reference to a more detailed report outlining the methodology of the study.

Interactive dashboard: This is relevant for subject matter experts to further explore and investigate different relations in the data. The dashboard should show all its features on one screen without overloading the user with too much information. Provide the scope and limitation, and ensure to provide some relevant and useful "buttons", "boxes" and "sliders".

Understanding the data value-chain: From collection to decision-making

ASEAN AUSTRALIA SMART CITIES TRUST FUND
Asian Development Bank

Interpretation, information & insights and decision-making

Regardless of the selected reporting type, it's important to provide the reader and decision-maker with the following contextual information about your data analytic study: 1) The purpose, scope and questions/objectives of your data analytic study. 2) Applied data processing and wrangling techniques, and ensure to be transparent about the validity and reliability of your data – especially specifying the limitation and application scope of your data and results.

The above information is important to keep in mind when interpreting the results to get insights for use in decision-making. Especially, for critical decisions related to citizens' safety and well-being, ensure to get enough evidence before making a conclusion. This can be achieved through triangulation, i.e., using different type of data collection approaches, different type of data and data sources or different type of data analytic approaches to expound on the topic of your investigation.

When presented to results from a data analytic study be critical. You can, for example, ask yourself the following questions:

1) Who has conducted the study – and is there a conflict-of-interest?
2) Is the results based on primary data sources (e.g., is the information about citizens background provided by themselves?) or on secondary data sources (e.g., is the information extracted from third-party or data brokers?).
3) What is the application scope and limitations of the data used in the study?
4) What is excluded in the data – and why?
5) Does other studies provide similar conclusion?

EXAMPLE

REFLECT & SUBMIT

REQUIRED SUBMISSION

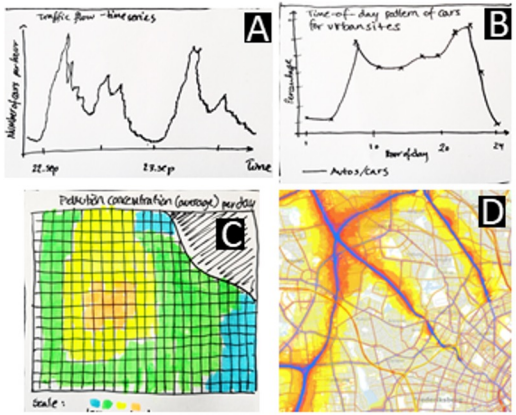
TIME: 20 minutes

REQUIRED MATERIALS: None

Show your results

Sketch some useful graphs and set-up for how you want to present your results and provide some explanation to what "story" you want to highlight with the specific graph

Type and sketch your answer here…



Example:

Descriptive analysis using time-series plot of the traffic flow (number of cars per hour) for a specific time period and location (Figure A). The figure shows that the traffic flow varies throughout the day - highest in the morning and afternoon. This suggest that "hour of day" is an important feature to include in our investigation. But we are not sure whether this pattern is prevalent in our data. Consequently, we investigate if that is the case in the next section.

We take the average of six months of data for urban sites (Figure B). The data is for the number of cars per hour relative to the total number of cars per day, shown as percentage. We see that a typical time-of-day pattern of cars – highest in the morning and afternoon. This confirms our previous observation.

We show the spatial distribution of air pollution concentration (average value for a day) as to visualise in which part of the city the pollution is highest and lowest (Figure C). We see that air pollution is largest in the centre of the city and in the northern part. This information can help assess which areas are most impacted by pollution.
Figure D shows the pollution applied on a map of the city. This helps to visualise the junctions and roads who are heavily affected by pollution affecting the surrounding areas.

EXERCISE SUBMISSION

REQUIRED SUBMISSION

Finished with this exercise?

Fill in the information below and follow the steps to submit you/your group's work:

Name of Individual/ Group members:	Type your answer here…
City:	Type your answer here…
Country:	Type your answer here…

STEP 1: Finalize this file by adding your city and last name onto this PPT's file name using the following format:

GLP Exercise 3_[City]_[Your Last Name]
(Example: **GLP Exercise 3_Baguio_Lucero**)

STEP 2: Go to the GLP Google Drive and save your renamed file inside the folder named after your city.

NOTE: We also recommend that you also save an offline copy of this file on your computer in case of file syncing issues.

If you encounter any issues with uploading your work, please contact your country's AASCTF national focal point.

< Back Next >

Thank you for finalizing Home Exercise 3

Congratulations on finalizing the last home exercise in the ASEAN Australia Smart Cities Trust Fund's Guided Learning Programme on urban data.

Please review your answers in preparation for the group sessions with the Ramboll team.

If you have any questions, please reach out to the National Focal Point

ASEAN AUSTRALIA SMART CITIES TRUST FUND
Asian Development Bank

Australian Government
Department of Foreign Affairs and Trade

ADB