|  |  |
| --- | --- |
| **Title** | **Data collection and analysis** |
| **Keywords (meta tag)** | **Data Analysis, Survey, Data Visualization, database** |
| **Provided by** | **University of Salento** |
| **Language** | **English** |
| **Area** | ***Entrepreneurship section*** |
| **Objectives / goals / learning outcomes** | |
| * Identify a potential Market by exploring the online databases and previous researches. * Sample potential customers * Define tools to collect data using the theory to develop appropriate and reliable questionnaires. * Analyze and interpret data with a friendly (and yet very powerful) approach based on Excel and Pivot Tables. | |
| **Description** | |
| This module will be divided into two main activities: first, you will learn how to evaluate needs and opportunities from the territory; the second activity is devoted to show the quantitative approach to data analysis; there, you will learn to extract knowledge from data collected from both the territory and potential customers.   * The identification of the potential market and customer is introduced by showing some of the principal and official open databases devoted to cultural tourism and entrepreneurship in the European context. An example on how to query and extract data from the database on virtual tourism observatory is proposed. The customer survey is introduced by giving the theory and procedures to select a representative sample of potential customers, in a way to understand the psychological profiles. * In the second part will be introduced the methods and tools to create and validate reliable and valid questionnaires. Moreover, to show how analyse data, we propose the use Excel with a very friendly approach based on Pivot Tables, which allows to explore one single variable and also to cross two or more variables in order to evaluate the relationships among different aspects. Moreover, the tool can create fashionable charts to improve the level of interpretation and visualization of the data collected. | |
| **Contents arranged in 3 levels** | |
| **Module name:**  **Data collection and analysis**  ***Introduction***  When a young ventures into the entrepreneurial journey, the starting point for success certainly is the territory. As a ready and existing knowledge base, the territory is made up of what works and what not, of particular needs and, at the same time, of hidden and explicit potentialities. The territory can tell a young entrepreneur whether there is room and support for his/her idea, whether there is a good chance for his/her business to grow, whether the expected results are worth the effort. However, in order to get this kind of message, the territory should be properly decoded: this is the goal of the current module, where you will learn to identify a market, as well as to conduct a survey on potential customers, where complete information on them is missing. Finally, you will learn to process the data of interest.  ***1.* First Unit: *Decoding the territory: needs and opportunities***  Decoding the territory means analyzing the market in order to secure the future of a business idea. The potential size of a market defines market opportunities: therefore, knowing about the current customer base allows for safer choices, when starting a business. 3 elements are particularly important to be determined:   * Market size.   How many potential customers are available? Are they available always or in a particular season? How are they?   * Profitability.   Are potential customers willing and able to spend on the services of interest?   * Potential growth.   Are there signs, studies or sources the market size will grow, stay relatively static, or decrease?  Have you ever imagined all these informations are just a click away? And that they are very precise, reliable and updated?  They come from the so-called “official-sources”. They might be national and international bodies who produce government census, statistical data, cultural tourism surveys and reports; they might be social media, who define size and features of groups of interest, as well as websites devoted to market research. A characteristic element of data provided by official sources is that they concern a whole “population” (e.g. all the tourists registered in a date-time period, in a country). Such data are either stored in downloadable database or available online for direct consultation. The most important official sources for public cultural tourism/entrepreneurship data are:   * [**Virtual Tourism Observatory**](https://ec.europa.eu/growth/tools-databases/vto/) * [**Eurostat**](https://ec.europa.eu/eurostat/statistics-explained/index.php/Tourism_statistics) * [**European Travel Commission**](https://etc-corporate.org/?page=research-intelligence)   They represent the starting point of a larger information research that might help in getting a clearer picture of how large your customer base could be and what kind of sustainability you can expect for the future. Public databases include a great amount of information at different levels of granularity and in different forms: that is why it is important to know how to “query” them and how to look at data visual representation.  **1.1. The Virtual Tourism Observatory.**  The VTO aims to support policy makers and businesses to develop better strategies for a more competitive European tourism sector. Their website offers a first ready-for-consultation representation of data: it helps getting a first glimpse on what is going on in the sector.    The visualization options are customizable, as well as the level (either the global EU or the country level) to be examined. The different options include dynamic representations of    Occupancy Seasonality Expenditure    Employment Non-EU tourism presence Regional Data  The graphical representations (such as bars vs horizontal lines, differently colored points, different bars) allow comparisons and ease up data interpretation. At the same time, the possibility to set up options and the responsiveness on the mouse click allows to narrow everything down and visualize information of interest.  **1.2. The VTO and Eurostat databases**  The VTO website provides a [*Country Profile*](https://ec.europa.eu/growth/tools-databases/vto/country-fact-sheets)area. By clicking on it, you will have the possibility to customize the data of interest you want to collect. Data available in the VTO come from the [Eurostat database](https://ec.europa.eu/eurostat/data/database)[[1]](#footnote-1).  In the VTO *Country Profile* section, let us say we want to explore how our country of interest is positioned in the European context. We might do so by comparing our country of interest with the European union    The next step is to select up to 6 indicators of interest.    Then, the *Compare* button will display online data comparisons. By clicking on the *Export to…* buttons, your tables will be at hand for further investigation with powerful tools such as Microsoft Excel.    The Eurostat database is much larger, therefore it needs a more focused research: the database page, in fact, concerns more than just information about tourism. Even though it might seem complicate to navigate these data, they carry information that might be crossed and looked at globally. The database area allows you to look at data by themes and timespan    Once you get to a particular table of interest, you have 3 options: navigate in the data browser through a data interface, or directly download the whole table.    As you refine (and redefine) your research, you might want to focus on a particular area of a region or city: in fact, tourism patterns can vary largely across regions (especially in big countries such as Spain or Italy).  The Eurostat homepage provides access to local data as well:    Within the *Statistics by* theme section, general *and regional statistics* offer data at a more granular level. Several tourism indicators will be available based on the NUTS (*Nomenclature of Territorial Units*) classification. Questions such as:   * Along with the touristic traffic, what kind of environment can one expect to find? * How is the quality of life? * How do transportation work?   will find proper answers, here. Dynamic data representation, as well as customizable download and data information are available.  **1.3. Additional Resources: TIPS and TRICKS**  Data are certainly useful. However, contextualizing them into your own entrepreneurial idea is what will boost their informative power.  When consulting data, either on graphical or tabular form, the recommendation is to progress through filtering questions: the starting point might be very generic (e.g. what is the sector trend over the years? How does it look like in the EU at a global level?); answer by answer, the questions get narrower, perhaps concerning the specific territory you want to create your idea in, or even comparing your territory with a more global level. As you get insight from the official data, you might want to know more about psychological features of your target customers. Sources such as social media (e.g. [*Facebook Audience Insights*](https://www.facebook.com/ads/audience-insights/), a free tool available on Facebook) might help you: if you’re interested in a particular region where Roman Routes in your entrepreneurial idea are located, you might search that area, define the features of the target customers you have in mind, verify their presence and interests; or you can even look at potential customers starting from their interests, and then reconsider/reframe your idea in light of the global data insight. The VTO, as well as the European Travel Commission, also direct towards official reports and surveys. Reports and surveys might provide insightful, focused and more qualitative information.  The examination and integration of different sources substantially increases your awareness of the territory, allowing you to identify your market size, the profitability of your idea and its potential of growth.  **2. Second Unit: *Data gathering and processing methodology***[[2]](#footnote-2)  Once you got insight from official data and identified your potential customers and subjects of interest, you might want to investigate the territory at a more granular, specific level. When you enter this level, you often find out that no data are available from the official sources/bodies. No worries! There is still a possibility to carry out an investigation on your own... If you know how to do it! In fact, the process of conducting a survey must be guided by precise criteria, as one might be limited with respect to the official bodies (running a census on a global population is often expensive and time consuming).  Prior to any concrete investigation, you must have a clear idea of your reference framework. Let yourself be guided by *the 5 W and the H*:   * *Who* is your reference population (e.g. potential customers)? * *What* is the area/topic you want to investigate (e.g. a particular kind of cultural tourism? Focused on sport activity rather than on typical food?) * *When* (e.g. time period of inquiry)? * *Where*? * *Why* (e.g. to inquiry how prone are people towards the idea, to understand strong points as well as hindrances to your idea)?   Once this information is in your mind, it’s time to take a look on:   * *How* to investigate,   that is, to know about data collection and processing techniques.    **2.1. The sampling**  Sampling represents a fundamental strategy: allows one to estimate the population parameters/results/perceptions by leveraging part of it. Sampling consists of extracting units from the population according to criteria that help generalize findings. In other words, a sound sampling strategy gives the possibility to state it is likely that a specific kind of customer would behave and perceive in a given way, based on the results obtained on part of them. However, generalizability depends on the sampling method itself. Sampling criteria might be divided into:   1. **Probabilistic,** where every element has a known nonzero probability of being sampled. Probabilistic sampling also involves a random selection at some point. In any probabilistic sampling method, the starting point is a list of the whole population. Extracting your customers of interest from a list of all the possible tourists registered in the summer season would allow you to generalize your conclusions.   Known probabilistic sampling strategies include:   * 1. ***Simple Random Sampling***: all the elements under investigation have the same probability of being part of the sample. Starting from a list of the whole population, the units are sampled randomly.      * 1. ***Systematic Sampling***: the study population according is ordered and, after a random start, elements are selected at regular intervals through that ordered list.  1. **Non Probabilistic**, where some elements of the population have no chance of selection (sometimes the latter are referred to as 'out of coverage'/'undercovered'), the probability of selection cannot be accurately determined. Therefore, they allow one to hypothesize rather than to generalize. Even though the evident shortcomings of this strategy, it can still be very useful when there is no knowledge about a certain phenomenon, as well as when a list of a whole population of interest is not available. Non probabilistic sampling strategies include:    1. ***Convenience Sampling***: the sample is taken from a group of people easy to contact or to reach;    2. ***Snowball Sampling***: after finding a group of initial respondents, these are used to recruit more respondents;   **2.2. Data Collection Techniques**  After determining the sampling criteria, it is time for the definition of the data collection tools: there is a very broad range of data collection tools, differing by their degree of structuredness (e.g. interviews move without a specific structure, living respondents free to develop their responses, while questionnaires are more rigorous and ask for shorter, defined answers). The Internet is a powerful source to find out whether somebody else already developed a data collection tool (such as a questionnaire) which is valid, reliable, appropriate to the sector you want to investigate, and perhaps… Directly downloadable! If you won’t find out any already existing tool, you might create one. However, it is important to keep some criteria in mind here, as well. Specifically, a questionnaire is a tool designed to collect information about aspects of interest (variables). 3 are the main steps of a questionnaire construction:   * ***Conceptual design***. If you detailed the previously mentioned 5 Ws and the H, you already set up a conceptual design for your survey; * ***Set up*** the questionnaire, that is:  |  |  | | --- | --- | | **What kind of information will be collected?**  **(Content)** | **How?**  **(Form)** | | Socio-Anagraphic information;  Attitudes;  Behaviors; | Open-ended question;  Closed-ended questions; |   Both forms of collecting information respectively have advantages and disadvantages, as you can guess by considering the table below:   |  |  |  | | --- | --- | --- | |  | **ADVANTAGES** | **DISADVANTAGES** | | **OPEN-ENDED QUESTIONS** | Freedom of expression, spontaneity.  Useful when it is not possible to really anticipate possible answers.  Useful for dealing with complex/delicate problems. | Too vague and difficult to understand  Coding issues (e.g. generic or inaccurate answers)  Answer quality depending on education level  Individuals who are not used to conceptualized in written form penalized  Demanding, high rejection rate | | **CLOSED-ENDED QUESTIONS** | Standardized (easy comparison).  Easy coding.  Multiple-answers help in clarifying the questions.  Elicit truthful answers when it comes to sensitive data.  Stimulate memory when focused on past events.  The interviewee is facilitated. | The interviewee could provide random answers.  The answers proposed may influence the response by excluding further alternatives.  In long inquiries, the order of questions may influence respondents. |   The question form also concerns formulation and order.   * ***Formulation***: when you build up a questionnaire, use   Simple Language (avoid aulic/flowery language);  Simple Syntax (avoid double negative, avoid requiring cognitive effort to the respondents);  Simple Content (investigate one feature at time, therefore avoid multiple statements in the same question).   * ***Order***:   Easiest answers first;  Follow a logical order;  Open-ended/sensitive questions at the end;  Alternate length and type.   * ***Verification***: from one side, it is important to evaluate the congruity between the measurement tool, as it has been prepared, and the cognitive needs of the survey; also, its functionality as a communication tool and as a useful tool for the interviewer. Verification is usually carried out through a pilot study, where the questionnaire is first administered to a reasoned sample. The final aim would be to allow the tool to produce generalizable results across groups and methods.   In practice, these aims are not always reachable. What matters is to keep in mind (and be aware) of the limits and the restrictions of the conclusions of the survey. It is desirable to collect data in particular formats: a table, where the elements of your data are separated by tabulation, comma or semicolon (.txt, .csv) or an Excel form (.xlsx). Once opened with Microsoft Excel, data processing can start.  **2.3 Data Analysis by using Pivot Table in Excel**  Pivot tables are interactive tables that allow the user to group and summarize large amounts of data in a more concise format. Pivot tables are built from a list of data by using some classical formulas in Excel, like sum, mean, min, max. In this tutorial we will use a simulated survey, designed as if the data were collected in the Apulia Region.    In order to create the pivot table, the spreadsheet with database must have the following characteristics:   * Having a column with duplicate values (e.g. Gender, Family, District, RomanRoutes). * Contain numerical values (e.g. age) to compare or add. Otherwise the only possible statistic is the count (e.g. years of residence).        1. Click on Insert Tab and then   click on Pivot Table   1. Excel will select all your data, but you can change the selection. 2. Create the Pivot Table on a new worksheet. 3. Click OK button to create your Pivot Table.   The Fields of a Pivot Table are mainly 3:  ***Rows***, ***Columns*** and ***Values***.  An extra field is used to define filters. To obtain the mean age of the interviewed considering Level of Education and whether he/she is close to a Roman Route, drag the fields in following boxes:   * Education in the *Rows Box* * RomanRoutes in the *Columns Box* * Age in *Values Box*, then select *Average*   Here the results.    Since there are no big differences in terms of mean age among the groups defined by using Education and RomanRoutes, it could be interesting to evaluate the amount of people that live near a Roman Routes.  You can do that by clicking on  in the *Values* box, then modifying its *Field Settings*. Here, you can select *Count*.    The table obtained tells us the number of people that leave near to a Roman Route with information about the level of education.  This could help to decide the type of business or sector. For instance cultural business, entertainment, food.    It is possible to **add filters**. Drag in the box Filter the variable **District**. You will obtain an extra window that allows you to select (or deselect) the statistical units in District of the Apulia Region. As it can be noted below, it is possible to **add extra fields**. Drag in the box Rows (you can drag as well in the box columns) the variable **Gender**.  This option will give extra information about the data, by going in deep in the gender differences of the sample analyzed.    It is possible to **create a Pivot Chart**. To insert the chart, follow these steps:  Select any cell of the Pivot Table. In the *Insert* tab click on Pivot Chart button.    You can **add filters** just by following the previous step for the Pivot Table. By adding the filter to the Pivot table you will add filter to the Chart.    The coolest thing about the Pivot table is certainly its dynamicity! Dynamic means it may get automatically updated. Let’s say you may need to add further data (perhaps, more recent data) to your table: you will add rows (that is, observations) to your raw table (the one containing the original, individual data). After done that, all you have to do is to *Change the Data Source* (by extending the selected area to the rows added).    These Excel functionalities spare you lots of manual work! In fact, once you update the *Data Source*, both the tabular and the graphical summaries will be automatically updated. Getting meaningful insight on how to make the most out of the Roman Routes nearby your area has never been so easy! | |
| **Socrative Questions (n.3):** | |
| 1. Why is decoding the territory so important for a tourism business idea? 2. What are the main Official Sources for data retrieving about cultural tourism and entrepreneurship? 3. How to query official databases? 4. How can the data insights be complemented by infos on additional sources? What do they say about potential customers and the market? 5. What to do when no detailed infos about target market and customers are missing? 6. What is sampling and how to run it? 7. How to build up a questionnaire? 8. How could one use their own data to produce summaries, interactive (pivot) tables and graphs? | |
| **5 glossary entries** | |
| **Official Sources**: official sources are reliable, credible resources based on a number of factors. It is a body of work created by an authoritative body in its field, which is widely recognized and considered as expert.  **Database**: a database is an organized collection of data, generally stored and accessed electronically from a computer system. Databases belonging to *Official Sources* are available online to consult and download.  **NUTS**: The Nomenclature of Territorial Units for Statistics (NUTS) is a hierarchical system dividing up the economic territory of the EU and the UK for the purpose of:   * Collection, development and harmonization of European regional statistics * Socio-economic analyses of the regions.   The Territorial units are divided into:   * NUTS 1: major socio-economic regions * NUTS 2: basic regions for the application of regional policies * NUTS 3: small regions for specific diagnoses   **Table**: traditional format of a collection of data. Rows usually indicate observations (e.g. individuals, customers, countries, etc.), while columns usually indicate variables (e.g. the phenomenon of interest, such as interests, market habits, features, etc). Sometimes, data is not perfectly organized and ready to be processed: let’s think, for example, of a data table retrieved from an official source, where columns display each a different year, while rows represent countries. In order to facilitate the graphical data representation, data must be transformed in order for columns to represent one single variable (in this case, all the columns belonging each to a different year should be pivoted into one column called Year, containing them all as rows): this might require data transformation operations (pivoting the table into long or wide format).  **Graph (or plot)**: a graphical representation of data. The type of graphical representation depends on the type of variable represented. Counts and percentages are usually represented with either bar graphs or pie charts, whereas counts of data-into-classes are represented by histograms. Scores are instead represented by points and connected points, while groups are rather represented by some graph aesthetic (e.g. color, shape, point dimension). | |
| **Bibliography and Further References** | |
| **Celine Roque. How to Define, Analyze, & Seize a Market Opportunity.**  [**https://business.tutsplus.com/tutorials/define-analyze-a-market-opportunity--cms-31875**](https://business.tutsplus.com/tutorials/define-analyze-a-market-opportunity--cms-31875)  **Corbetta, P. (2003). *Social research: Theory, methods and techniques*. Sage.**  **Excel Easy: #1 Excel Tutorial on the net.**  [**https://www.excel-easy.com/**](https://www.excel-easy.com/)  **TutorialsPoint - Simply & Easy Learning. Learn Statistics.** [**https://www.tutorialspoint.com/statistics/index.htm**](https://www.tutorialspoint.com/statistics/index.htm)  **Yaroslav Lehenchuk. How To Research The Market And Identify Opportunities.**  [**https://producttribe.com/marketing-amp-partnerships/market-research-guide**](https://producttribe.com/marketing-amp-partnerships/market-research-guide) | |
| **Related Material** | ARLR\_ Training\_fiche\_DC&A\_revised.docx |
| **Related PPT** | ARLR\_ Training\_fiche\_DC&A\_revised.ppt |
| **Reference Link** |  |

1. Eurostat (European Statistical Office) is a Directorate-General of the European Commission. It provides statistical information to the institutions of the European Union (EU) and promotes the harmonisation of statistical methods across Europe. A larger and broader data collection is available on its website. [↑](#footnote-ref-1)
2. For further details, see Corbetta (2003). Here, provided further details about *[sampling design](https://www.tutorialspoint.com/statistics/sample_planning.htm)* and [*sampling strategies*](https://www.tutorialspoint.com/statistics/sampling_methods.htm). [↑](#footnote-ref-2)