International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Data Preparation in Context of Social Sciences Research

Satwik Sahay Bisarya[1], Anjali Shukla[2], and Santosh Kumar[3]

[1]Assosiate Professor, Faculty of Agriculture Sciense and Technology, Madhyanchal Professional University, Bhopal (M.P.)
[2]Assistant Professor, Institute of Agriculture Science, SAGE, Indore (M.P.)
[3]Assistant Professor, Faculty of Agriculture Sciense and Technology, AKS University, Satna (M.P.)

ABSTRACT
In many research fields, including social sciences one needs to prepare quality data by pre-processing the raw data. An essential step in the data analysis process is data preparation. The raw data which we collect, never comes in a form that can be used immediately. We have to take data and draw some useful information by compiling data into right format without errors. Data preparation can be broadly classified into three phases, extract, transform and load. While extracting we are focused on connecting and joining data. In transforming stage, we do some simple auditing to see what we are working with and finding the loopholes in the data. In last phase which is loading, we ultimately end up with the complete data set ready for analysis in tableau. In this book chapter, we have tried to explain the various steps involved in data preparation beginning from questionnaire checking to storage of data.

Keywords: Data Preparation, Data Cleaning, Outliers, Tabulation, Storage

Introduction

In many research fields, including social sciences one needs to prepare quality data by pre-processing the raw data. An essential step in the data analysis process is data preparation. Several organizations or businesses are interested in finding ways to turn the data into cleaned forms that may be used for highly profitable reasons, even if there is a lot of low-quality information available in a variety of data sources and on the Internet. Data analysis with the intention of cleaning the raw data is urgently required to achieve this goal. The raw data which we collect (through any tools like questionnaire, direct interview, online survey), never comes in a form that can be used immediately. We have to take data and draw some useful information by compiling data into right format without errors.

Before data preparation we should keep asking some questions to ourselves in mind

- What is the condition of data (source)?
- Where the data generated is going to be used?
- Whom we are supplying information?
- How we can improve the data?
- How much time and effort will be required to process the data into useful information?

Basically, data preparation can be divided into 3 broad phases:

Fig.1 Phases of data preparation

I. **Extract:** In this phase we are focused on connecting and joining data. The data can be spread over different files or sheets so we assemble them in one place.
II. **Transform:** In this phase, we do some simple auditing to see what we are working with, finding the loopholes in the data, changes to be made are noted, shaping the data, cleaning the data also, deciding which part of data to come as final output to the end user.
III. **Load:** In this phase we ultimately end up with the complete data set ready for analysis in tableau.

Steps to be followed while preparing the data

1. Questionnaire checking
2. Editing
3. Coding
4. Classification
5. Tabulation
6. Graphical representation
7. Data cleaning
8. Data adjusting
9. Storage of data

1. Questionnaire checking

If the questionnaire has some errors or the questions are asked in improper way, or wrong scales are used than the sole purpose of the survey may not be fulfilled. The questionnaire can be checked in three parts for precision and accuracy:

a. Outline & the format: It includes the introduction part, demographic information and arrangement of the questions. The introduction part includes the title of the study, purpose of the study, duration of survey. The confidentiality of the respondents should be maintained while conducting the research. Demographic information includes the respondent's name, age, gender and contact number if provided by the respondent.

While preparing format of the questions, concise, clear, and appropriate questions with answer choices should be framed so that answering questions won't confuse the respondents and they can clearly provide data on their experience. For this, first questions should be close ended, topic oriented and arranged from general to specific. If any rating scales are used, should be written before the question. Sensitive questions should be asked tactfully & should be placed at last. At the end of the questionnaire, any expression of gratitude towards the respondent should be given for giving their valuable time. Information regarding how data will be used and the survey results could be provided.

b. **Questions**: It should be simple & concise and should not contain difficult terminologies, acronyms, or jargons unfamiliar to the respondent. Initial questions are very much important to verify whether the person is willing to continue the survey or not so they should be framed with utmost care. All possible answers to a question should be mentioned in the options or *"Other"* option should be included. Sensitive questions should have *"prefer not to answer"* option.

c. **The pre-test:** The questionnaire may be sent to friends, colleagues, experts to check the possible errors before actual use. Responding time should be noted down to determine the time required to complete the survey.

2. Editing

It is the process of reviewing and adjusting the collected data. Editing enables us to detect and omit any types of error to make data complete, consistent and accurate. Unreadable, unclear, incomplete, unanswered responses are completed through editing. It involves thorough inspection of the completed questionnaires.

Editing can be completed at two stages:

a) **Field editing:** It is done then and there at the time of interviewing the respondent or as soon as possible by the investigator in abbreviated form. Care should be taken while filling the missing data that it should not be a wild guess how the respondent would have answered to the question.

b) **Central editing:** It is done when all the questionnaires have been completed and returned to the investigator at his workplace. Editing can be done by a single editor in case of small field study or by group editors in case of large field study.

The corrector may correct the obvious mistakes like missing gender, or writing correct information written at wrong place by revising the other information documented in the schedule. If necessary, the respondent may be communicated for description. The date of the editing may be documented for future references.

3. Coding

It is the process of allocating alpha-numeric symbols to the answers for easy comprehension to the investigator. The responses are then divided into different classes of suitable sizes, exhaustive and mutually exclusive of each other. Coding is necessary for efficient data analysis. It makes the tabulation of data easier for further analysis.

For example: for monthly income of respondent

a) Less than Rs.5000
b) Rs.5000-10000
c) Rs.10000-15000
d) Rs.15000-20000
e) More than Rs.20000

We can code the class *"Less than Rs.5000"* as 1, *"Rs.5000-10000"* as 2, *"Rs.10000-15000"* as 3, *"Rs.15000-20000"* as 4 and *"More than Rs.20000"* as 5.

4. Classification

It is the method of arranging the data into identical classes according to the common features present in the data or sorting out them into different but related part. The volume of data collected is reduced by classifying them which facilitates easy comparison, study the relationship and helps in statistical treatment of data.

Classification can be of various types like:

a) Classification according to attributes

For the data which cannot be measured quantitatively like sex, caste, education. It is known as qualitative classification.

b) Classification according numerical characteristics

When the data is having some measurable characteristics like height, weight, sales, profit, and income, etc. quantitative classification is followed.

c) Classification according to geographical location

When the data is classified according to geographical locations like village, block, district, state etc., the classification is known as spatial classification or geographical classification.

d) Classification according to time

When the data is classified on the basis of time, it is known as chronological or temporal classification. In this, data are classified either in ascending or in descending order with reference to time such as week, months, year, etc.

5. Tabulation

It is the systematic or logical organization of figures into rows and columns which facilitates easy comparison and makes statistical analysis time efficient. Tables help in summarization and compression of data. It presents facts in minimum possible space. The tabulation may be simple (one-way tables related to one characteristic of data) or complex (two-way or multi-way tables, which gives information about several interrelated characteristics). Every table should be given their title and distinct number to facilitate easy reference. Units of measurements should be given in the tables. Explanatory footnotes may be given below the table with reference symbol.

6. Graphical Representation

It is visual display of data and statistical results that preserves the characteristics of data and displays them at a glance. It is an effective tool for clear understanding and interpretation of collected data. Data can be represented in several ways such as bar graph pie chart, line graph, pictograph, histogram, frequency distribution, scatter plot etc.

a. **Bar graph:** It is a graphical representation of data in which bars of uniform width are drawn with identical gaps or spacing between them. The values of the data are represented by the height of each bar.

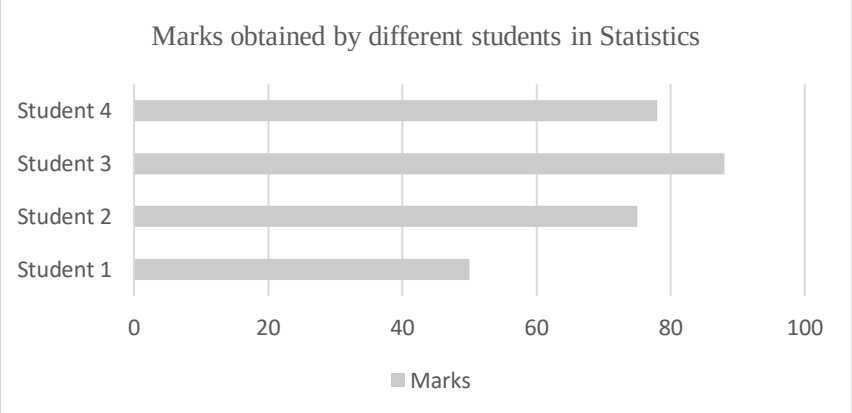

Fig.2 Representation of marks through bar graph.

b. **Pie Chart:** A pie chart, sometimes known as a circle chart, is a circular statistical graphical representation that shows numerical proportion through slices of data. Each slice's arc length in a pie chart is correlated with the number it shows.

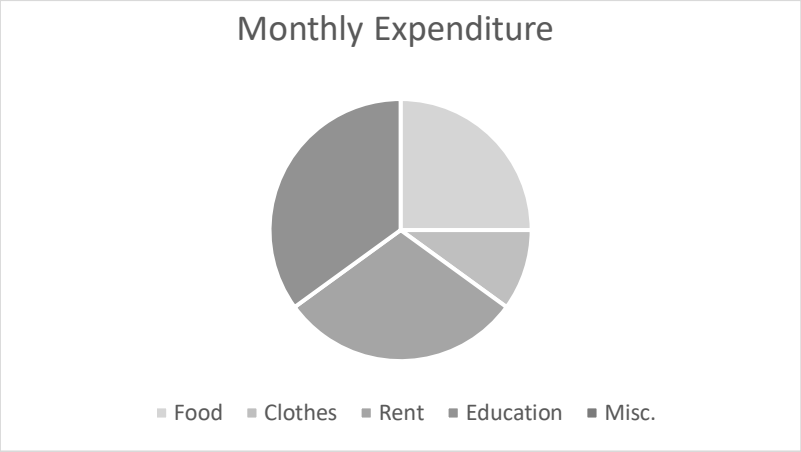

Fig.3 Representation of monthly expenditure through pie chart.

c. **Line Graph:** A graph that uses points and lines to depict change over time is known as a line graph. It is also known as line chart or line plot. A line connecting many points or a line illustrating the relationship between the points is shown on the graph. It combines a series of subsequent data points with a line or curve to illustrate quantitative data between two changing variables. These two variables are compared in a vertical axis and a horizontal axis on a linear graph.

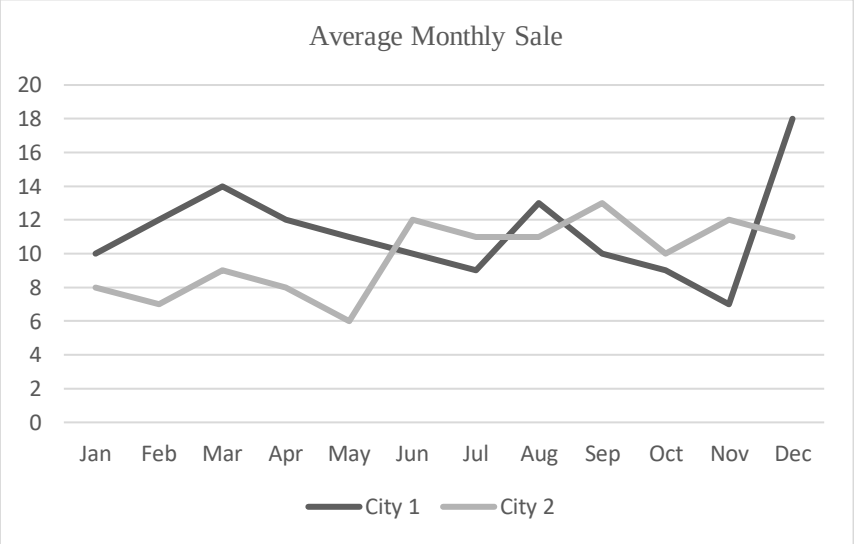

Fig.4 Representation of average monthly sale through line chart.

d. **Pictograph:** Pictographs are charts which uses graphics and icons that are appropriate to the data to represent the data. A pictograph frequently has a key which describes what each emblem or image stands for. All the icons of pictogram must have the same size, but we can use a fraction of an icon to represent each fraction of the total.

Number of Flowers Sold	
Monday	✿ ✿ ✿ ✿
Tuesday	✿ ✿ ✿ ✿ ✿ ✿
Wednesday	✿ ✿
Thursday	✿ ✿ ✿
Friday	✿ ✿ ✿ ✿ ✿
Saturday	✿ ✿ ✿ ✿ ✿ ✿ ✿
Sunday	✿ ✿ ✿ ✿
Legend: ✿ = 10 Flowers	

Fig.5 Representation of number of flowers sold through pictograph.

e. **Histogram:** A histogram is a graphic depiction of a frequency distribution with continuous classes that has been grouped. A series of rectangles with bases equal to the distances between class boundaries and areas proportional to frequencies in the associated classes make up the area diagram. It is similar to that of bar graph but it doesn't have space between the bars.

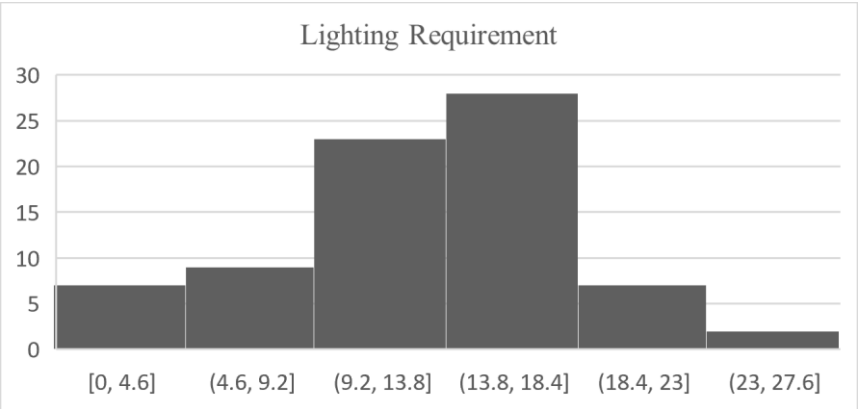

Fig.6 Representation of data through histogram.

f. **Frequency distribution:** The number of times a value appears in a dataset is its frequency. The pattern of frequencies for a variable is known as a frequency distribution. The frequency with which each potential value of a variable appears in a dataset.

Age (in years)	Frequency
Less than 25	60
25-50	30
More than 50	10

Fig.7 Representation of data through frequency table.

g. **Scatter plot:** It is a type of data representation that shows the relationship between different variables by placing numerous data points between an x- and y-axis. Each of these data points looks dispersed around the graph, giving this type of data representation its name. Scatter plots can also be known as scatter diagrams. They are used to determine whether there are patterns or correlations between two variables is present or not.

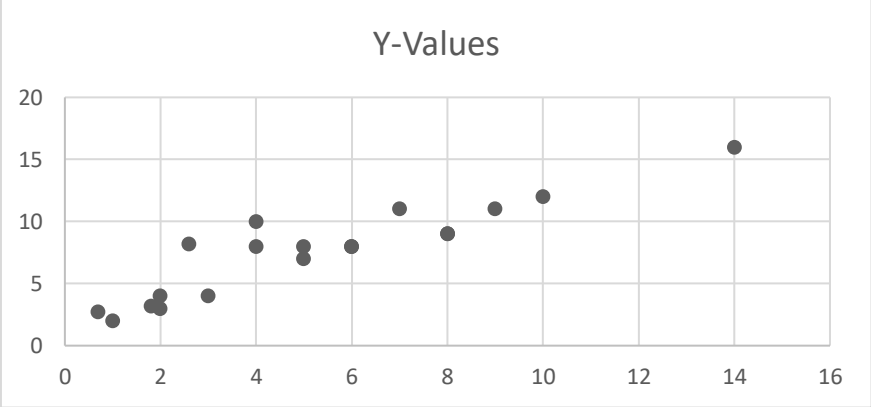

Fig.7 Representation of data through scatter plot.

7. Data Cleaning

It is the process of detecting and eliminating incorrect, wrongly formatted, duplicate and missing records. Cleaning also deals with the consistency of data. It is the process of eliminating some data which are not needed from entire data. Some common steps of cleaning involve removing unwanted observations, removing punctuations, extra spaces, and numbers. It also fixes structural errors (errors arise during measurement, data transfer, etc. and filtering records which we don't want to keep (outliers) from source data.

Data purification: In this stage we check for the missing data, outliers, normality and linearity of data.

Null values (Missing data): Dealing with the null values, null values do not mean "0", it can be missing, invalid or unknown data at the time of record. For missing data, we can use mean substitution, Hot or cold deck method or regression method.

Outliers: It is the abnormal observation from the normal values in the random sample of a population. Outliers should be examined cautiously. Before excluding them, the investigator should consider from where they appeared. Sometimes outliers may contain important information. Outliers can be checked through various methods like:

Visual Tools

a) Box plot construction

b) Scatter plot construction

With Mathematical tools:

a) Z score method

b) IQR score method

Normality: The suppositions of normality must be checked to apply various parametric tests (correlation, regression, t-test, ANOVA) as the validity of the test is subjected to the normality of data. When assumptions of normality do not hold, it is challenging to draw precise and consistent conclusions. Test for normality includes,

a) Visual methods: Histogram, stem-and-leaf plot, boxplot, P-P plot (probability-probability plot), and Q-Q plot (quantile-quantile plot) are used for checking normality visually.

b) Normality test:

- Kolmogorov-Smirnov (K-S) test
- Lilliefors corrected K-S test
- Anderson-Darling test
- Anscombe-Glynn kurtosis test

c) Test of Normality using SPSS

Skewness: It is the measure of symmetry, when the data is not following Normal Distribution, it can be positively or negatively skewed.

Kurtosis: It is the measure of flatness or peakedness of distribution. A distribution can be leptokurtic, mesokurtic or platykurtic.

In a normally distributed data, the value of skewness and kurtosis both are zero. Significant skewness or kurtosis undoubtedly specifies that the data is not normal. It is may be due to the presence of outliers in the data. Skewed and kurtotic data can be corrected by various transformations or by using software to make it normal.

Linearity of data: Linearity of data should also be checked as many statistical methods require an assumption of linearity of data. Generally, a normally distributed data also follows linearity but not always. The data are closed to the regression line if it follows linearity. Linearity can be easily checked by using SPSS.

8. Data Adjusting

Sometimes data adjustment is done to make it statistically correct and enrich the data quality. It can be done by several ways; some are mentioned below:

a. By giving weight to variables: Data is adjusted by giving weights to each respondent according to the comparative importance to others. It is done to increase or decrease the number of respondents in the sample so that the sample data is more symbolic to target population.

E.g., While calculating the weather forecast report through the past trends; recent years are given more weightage than previous ones.

b. Variable re-specification: It involves transforming the data to make new variables or modify existing variables. It done to create variables that is more consistent with the objective of the research.

E.g.: While asking the respondents about the smartphone usage on 7- point scale, we have 7 different response categories in our survey. So now we can collapse the 7 responses into 3-4 categories.

Response category (7-pt scale)	Modified response category	
1	Least likely to use	Those who selected 1,2,3
2		
3		
4	Neutral	Those who selected 4
5	Most likely to use	Those who selected 5,6,7
6		
7		

d. **Scale transformation:** It is the manipulation of scale to it comparable with other scales and to make the data prepared for comparison.

9. Storage of data

The prepared data are then stored for future analysis using suitable data analysis strategy. Storing data properly can save lots of time in finding and interpreting it during and after the research. To store the data properly, we need to choose the storage media wisely. All storage locations are not equally suitable, it varies according to the type of data we are having. Data can be stored using cloud services or in portable devices like CD, hard drive, USB flash drive. The raw data should be protected as it is the basis of all research. Several copies can be stored at various locations for easy retrieval.

Conclusion

As the research work progresses, more files are added which will create hassle while searching some particular documents. So proper naming and structure should be there of all versions of data.

References

1. Costello, & Blackshear, L. (2019). Prepare Your Data for Tableau: A Practical Guide to the Tableau Data Prep Tool. Apress L. P. Retrieved from https://doi.org/10.1007/978-1-4842-5497-4 on 10December 2022
2. Field, A. (2013). Discovering statistics using IBM SPSS statistics. sage.
3. Fred N. Kerlinger, “Foundations of Behavioral Research: Educational, Psychological and Sociological Enquiry” Wadsworth Publishing Company; 2nd Revised edition.
4. Hancock, G. R., Mueller, R. O., & Stapleton, L. M. (2010). The reviewer's guide to quantitative methods in the social sciences. Routledge.
5. Jürgens, P., Stark, B., & Magin, M. (2020). Two half-truths make a whole? On bias in self-reports and tracking data. Social Science Computer Review, 38(5), 600-615.
6. Kwak, S. K., & Kim, J. H. (2017). Statistical data preparation: management of missing values and outliers. Korean journal of anesthesiology, 70(4), 407-411.
7. Lane, J., Stodden, V., Bender, S., & Nissenbaum, H. (Eds.). (2014). Privacy, big data, and the public good: Frameworks for engagement. Cambridge University Press.
8. Lazer, D. M., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., ... & Wagner, C. (2020). Computational social science: Obstacles and opportunities. Science, 369(6507), 1060-1062.

9. Mneimneh, Z., Pasek, J., Singh, L., Best, R., Bode, L., Bruch, E.,& Wojcik, S. (2021). Data acquisition, sampling, and data preparation considerations for quantitative social science research using social media data.
10. Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics–Challenges in topic discovery, data collection, and data preparation. International journal of information management, 39, 156-168.
11. Zhang, S., Zhang, C., & Yang, Q. (2003). Data preparation for data mining. Applied artificial intelligence, 17(5-6), 375-381.