Batch: D2 **Roll No.:** 16010122323

Experiment 01

Grade: AA / AB / BB / BC / CC / CD /DD

Title: Data Collection and finalizing dataset from problem domain

---

Objective:

1. To learn how to collect the dataset

2. To learn sources of dataset

3. To asses the dataset based on Metrics to Measure Data Quality

4. To finalize the features of dataset

Course Outcome:

CO1 : Learn how to locate and download datasets, extract insights from that data and present their findings in a variety of different formats.

Books/ Journals/ Websites referred:

1. Data Visualization made simple New York: Routledge - Kristen Sosulski, First edition, 2019
2. Sosulski, K. Data Visualization Made Simple: Insights into Becoming Visual, First edition, 2018
3. https://www.kaggle.com/uciml/adult-census-income
4. https://archive.ics.uci.edu/ml/datasets/adult
5. https://ori.hhs.gov/education/products/n_illinois_u/datamanagement/dctopic.html
6. A review of research process, data collection and analysis - Surya Raj Niraula

Resources used:

1. https://www.kaggle.com/uciml/adult-census-income
2. https://archive.ics.uci.edu/ml/datasets/adult

Theory:

Data collection is the process of gathering and measuring information on the variables of interest is an established, systematic fashion. It is defined as the collecting, measuring and analyzing accurate data for research using standard validated techniques.

Types of data:
The data collected is segregated into two types:

1. Qualitative type:
   There are 3 subtypes in this type of data:
   - Nominal data:
     Values that don't possess any natural ordering.
     Eg: colour of mobile.
   - Ordinal data:
     Values that have natural ordering.
     Eg: Small< medium<Large.
   - Categorical data:
     Data that represents characteristics.
     Eg: Height.

2. Quantitative type:
   There are 2 subtypes in this type of data:
   - Discrete:
     Numerical values that fall under the integers or whole numbers.
     Eg: Cameras, bullets.
   - Continuous:
     Data used to represent measurements.

Some of the data collection methods include:

- Case Studies
- Checklists
- Interviews
- Observation sometimes,
- Surveys or Questionnaires
  - In-person
  - Mail
  - Phone
  - Web/online

The first step in gathering system data is to determine what data is required for building the model.
For eg:

Hand writing recognition
Sports classification
Human Identification
Automatic Attendance system

The different criteria used to measure data quality are:

- Accuracy: for whatever data described, it needs to be accurate.
- Relevancy: the data should meet the requirements for the intended use.
- Completeness: the data should not have missing values or miss data records.
- Timeliness: the data should be up to date.
- Consistency: the data should have the data format as expected and can be cross reference-able with the same results.

Metrics to measure data quality are:

Metric	Definition	How to calculate
Ratio of Data to Errors	How many errors do you have relative to the size of your data set?	Divide the total number of errors by the total number of items.
Number xof Empty Values	Empty values indicate information is missing from a data set.	Count the number of fields that are empty within a data set.
Data Transformation Error Rates	How many errors arise as you convert information into a different format?	How often does data fail to convert successfully?
Amounts of Dark Data	How much information is unusable due to data quality problems?	Look at how much of your data has data quality problems.
Email Bounce Rates	What percentage of recipients didn't receive your email because it went to the wrong address?	Divide the total number of emails that bounced by the total number of emails sent, then multiply by 100.
Data Storage Costs	How much does it cost to store your data?	What is your data storage provider charging you to store information?
Data Time-to-Value	How long does it take for your firm to get value from its information?	Decide what "value" means to your firm, then measure how long it takes to achieve that value.

Problem domain

The dataset we have chosen is - Adult Census Income (available in Kaggle and UCI repository). This data was extracted from the 1994 Census bureau database by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics). It involves using personal details such as education level to predict whether an individual will earn more or less than $50,000 per year.

AIM: The task at hand is to predict and determine whether a person makes over $50K a year or not.

Brain stormed features of Dataset\

This dataset has 15 attributes which provide a detailed overview of the people living and working in the USA.

The attributes can be listed as follows:

Categorical Attributes:

- Workclass: Individual's work category'
    - Self-emp-inc
    - Private
    - Self-emp-not-inc
    - Federal-gov
    - Local-gov
    - State-gov
    - Without-pay
    - Never-worked
- Education: Individual's highest education degree
    - Some-college
    - Bachelors
    - 11th
    - HS-grad
    - Prof-school
    - Assoc-acdm
    - Assoc-voc
    - 9th
    - 10th
    - 12th
    - Masters
    - Preschool
    - 5th to 8th
- Marital-status: Individual's marital status
    - Married-civ-spouse
    - Never married
    - Divorced
    - Separated
    - Widowed

    - Married-spouse-absent
    - Married-AF-spouse

- Occupation: Individual’s occupation
    - Tech-support
    - Craft-repair
    - Other-service
    - Sales
    - Prof-specialty
    - Handlers-cleaners
    - Machines-op-inspect
    - Adm-clerical
    - Farming-fishing
    - Transportation-moving
    - Armed-forces

- Relationship: Individual’s relation in a family
    - Wife
    - Own-child
    - Husband
    - Not-in-family
    - Other-relatives

- Race: Race of individual
    - White
    - Black
    - Asian-Pac-Islander
    - Amer-Indian-Eskimo
    - Other

- Sex:
    - Female
    - Male

- Native-Country: Individual’s native country:
    - India
    - Japan
    - Dominican-Republic
    - Laos
    - Ecuador
    - Taiwan

    - Thailand
    - Yugoslavia
    - El-Salvador
    - Italy
    - Poland
    - England
    - France
    - Mexico
    - Etc.

Continuous Attributes

- Age: Continuous. Age of an individual

- fnlwgt: Final weight, continuous.

  The weights on the CPS files are controlled to independent estimates of the civilian noninstitutional population of the US. These are prepared monthly for us by Population Division here at the Census Bureau.

- Capital-gain: continuous.
- Capital-loss: continuous.
- hours-per-week: continuous. Individual’s working hour per week.

WE WILL BE CHOOSING THE FOLLOWING ATTRIBUTES:

1. Workclass
2. Education
3. Marital Status
4. Occupation
5. Relationship
6. Race
7. Sex
8. Native Country
9. Age
10. Income

We chose the above-mentioned attributes as they give us the necessary information needed to solve the problem statement. The work class tells us about the kind of job a person does. Education, occupation and income can help in predicting the basic stand of an individual in the society. Marital status, relationship, race, age, sex and native country can help us identify any discrimination (if any) in payment which might affect the yearly income of an individual. These are key attributes that will help us get an idea to predict the income of an individual.

Motivation for the selected dataset

This experiment aims to predict whether the income of U.S. population exceeds $50K/yr or not based on census data provided by Census bureau database, while considering different factors such as age, work class, gender, marital status, education, race, occupation etc. using exploratory analysis and classification algorithms. The dataset contains 32561 records.

The dataset encourages to draw valuable insights and conclusions. The conclusions drawn might help in delivering wiser decisions. In addition to it, suggestions could be given based on the predictions to students who are in need to pursue higher education and people who are spending less time in the workplace.

Income is instrumental in deciding a person’s standard of living and the financial status in the society. It plays a key role in determining growth of nation. Our aim is to identify meaningful insights which can be the basis for many cleverer decisions. Our dataset contains 32561 records with various attributes such as occupation, age, relationship, hours per week, education, income and so on. Exploratory analysis will be done between dependent and independent variables.

Not all the attributes are relevant for our analysis. The selection of the useful attributes will be based on the outcomes of the various algorithms. The variables are numeric as well as multiple factors.

Source of dataset

We first came across this dataset on UCI repository and its also available on Kaggle. The csv file used for this project has been downloaded from Kaggle, links have been mentioned above.

Sample of Finalized dataset and its source

The sample excel file of the dataset can be accessed using this link:

https://docs.google.com/spreadsheets/d/1sq9emMpz8ASwvTGdMB1I1vfOhmRsiZjTz-rgGm9WgJg/edit?usp=sharing

Justification for choosing above dataset

Money is needed to survive in the world and this money is made available to common people through their income. We chose this data set to learn about the income distribution amongst the citizens if the USA. Census data includes attributes like race, occupation, daily income, etc. which give us an overview of the situation of an individual. This overview is necessary and needs to be studies thoroughly before making huge political discissions that affect everyone in the society. Income is instrumental in deciding a person’s standard of living and the financial status in the society. It plays a key role in determining growth of nation. Our aim is to identify meaningful insights which can be the basis for many cleverer decisions. Our dataset contains 32561 records with various attributes such as occupation, age, relationship, hours per week, education, income and so on. Exploratory analysis will be done between dependent and independent variables.

Thus, by analyzing this data set better public decisions can be made for the upliftment of those who have a lower income in the society.

Conclusion (Students should write in their own words):

Through this experiment, we were introduced to various data sets. We learnt how to analyze the data in a particular data set and how to properly segregate it and use it for solving topic related problems.

Post Lab Question:

1. **Explain Role of Data in the Application Design.**

- Data and design need to be integrated together to provide the enhanced user experience for your products. Design decisions based on data can never be challenged.
- You need to learn about how to collect data at different design stages, how to visualize this data, and how to analyze the visualizations so that you can use these findings to finalize your design.
- Data-driven design validates the concept of a user-centric design approach that enforces the involvement of users throughout the design process.
- Data-driven UI design can require a variety of different kinds of data to determine the best way to create an optimal user experience. This data can include things like website or app analytics on an existing iteration of a product, user interviews, A/B and multivariate test results, behavior flows, and other types of UX research.

2. **Write different types of Data with Example.**

There are 5 types of data:

1. Numerical

Numerical data is any data where data points are exact numbers. It is a measurable quantity. Numerical data can be characterized by
• continuous data can assume any value within a range(average score of student)
• discrete data has distinct values(no. of students enrolled)

2. Categorical

Categorical data represents characteristics. Categorical data can take numerical values. In the context of classification, categorical data would be the class label. This would also be something like if a person is a man or woman, or property is residential or commercial.

3. Ordinal Data

Ordinal data is a mix of numerical and categorical data. In ordinal data, the data still falls into categories, but those categories are ordered. An example is that we just take quantitative data, and splitting it into groups, so we have bins or categories of other types of data.

4. Timer Series Data

Time series data is a sequence of numbers collected at regular intervals over some period of time. It is very important, especially in particular fields like finance.

5. Text

Text data is basically just words. A lot of the time the first thing that you do with text is you turn it into numbers using formulation.

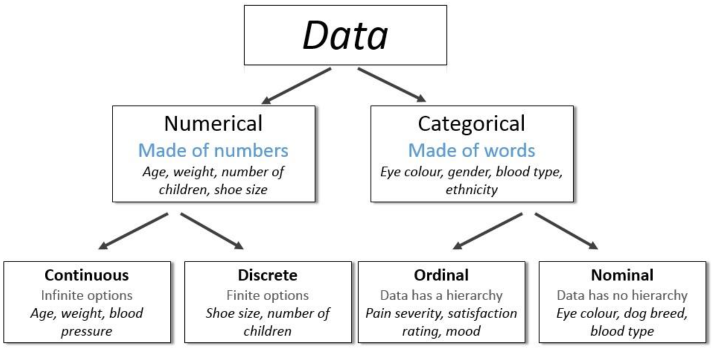
Data
Numerical
Made of numbers
Age, weight, number of children, shoe size
Categorical
Made of words
Eye colour, gender, blood type, ethnicity
Continuous
Infinite options
Age, weight, blood pressure
Discrete
Finite options
Shoe size, number of children
Ordinal
Data has a hierarchy
Pain severity, satisfaction rating, mood
Nominal
Data has no hierarchy
Eye colour, dog breed, blood type

Figure 1: Graphic showing Types of Data