Lecture Note 2: Data Types

1. (Tabular) Data

In the context of tabular data, a data table or spreadsheet serves as a fundamental structure for organizing information. A dataset may consist of multiple relevant data tables, each containing specific types of data related to the analysis.

Columns in the data table represent variables, which can either be attributes (also known as features) or targets. Attributes are characteristics or properties of the subjects being studied, while targets are the outcomes or values that we want to predict or analyze.

Rows in the data table represent individual records, observations, instances, or data points. Each row corresponds to a specific entry or data set for a particular subject or entity. For example, in a dataset about customers, each row would represent a unique customer, and the columns would contain their attributes (e.g., age, gender, income) and targets (e.g., purchase behavior, churn status).

1.1 Types of Data/Variables

Categorical/Qualitative Variables: Categorical variables represent characteristics or qualities that are not inherently numeric and can be divided into two main types: nominal and ordinal variables.

1. Nominal Variables: Nominal variables are categorical variables with no inherent order or ranking among their categories. Examples include colors, gender (male/female), or types of animals. They are typically represented using labels or names.
2. Ordinal Variables: Ordinal variables are categorical variables with a clear order or ranking among their categories. However, the intervals between categories are not necessarily equal. Examples include education levels (e.g., high school, bachelor's, master's) or customer satisfaction ratings (e.g., low, medium, high).

Numeric/Quantitative Variables: Numeric variables, also known as quantitative variables, are measurements represented by numbers and can be further divided into two types: discrete and continuous variables.

1. Discrete Variables: Discrete variables are numeric variables that can only take specific, separate values. These values are often integers and cannot be subdivided. Examples include the number of students in a class, the count of product defects, or the number of cars in a parking lot.
2. Continuous Variables: Continuous variables are numeric variables that can take any value within a certain range, including decimal numbers. They have infinite possible values, making them suitable for measurement on a continuous scale. Examples include height, weight, temperature, or sales revenue.

Understanding the different types of data or variables is crucial in data analysis and modeling, as it determines the appropriate statistical methods, visualizations, and interpretation techniques to be used for effective data exploration and insights extraction.

By organizing data in this tabular format, it becomes easier to analyze and process information, facilitating various data manipulation, visualization, and modeling tasks in the context of cross-sectional data analysis.

2. Datasets for the Course

Throughout the course, we will extensively utilize two datasets obtained from kaggle.com. The first dataset comprises house sale prices and related information from May 2014 to May 2015 in King County, WA (https://www.kaggle.com/harlfoxem/housesalesprediction). The second dataset includes details on over 10,000 credit card holders from a bank (https://www.kaggle.com/competitions/1056lab-credit-card-customer-churn-prediction/data). Below is the description of the data fields present in each dataset.

Table: Data Field Description - House Sales Dataset

Variable	Description
Id	Unique ID for each residential unit sold
Date	Date the residential area was sold
Price	Price of each residential unit sold
Bedrooms	Number of Bedrooms
Bathrooms	Number of Bathrooms where "half (1/2) bath" (or "powder room") containing just a toilet and sink; and "3/4 bath" containing toilet, sink, and shower.
Sqft_living	Square footage of the living room space
Sqft_lot	Square footage of the land space
Floors	Number of floors
Waterfront	A dummy variable for whether the residential unit was overlooking the waterfront or not.
View	An index from 0 to 4 of how good the view of the property was: 0 = No view, 1 = Fair 2 = Average, 3 = Good, 4 = Excellent
Condition	An index from 1 to 5 on the condition of the apartment: 1 = Poor- Worn out, 2 = Fair- Badly worn, 3 = Average, 4 = Good, 5= Very Good
Grade	1-14 Index where 1-3 falls short of Building construction and design. 7 is an average level. 11 to 13 implies high quality design level. 14 is the extraordinary.
Sqft_above	Interior housing space above ground level
Sqft_basement	Interior housing space below ground level
Yr_built	The year the residential unit was built
Yr_renovated	The year last innovation was recorded
Zipcode	Zip code area of the residential unit
Lat	Latitude
Long	Longitude

Sqft_living15	Square footage of interior living space in comparison with the nearest 15 neighbors
Sqft_lot15	Square footage of land lots space in comparison with the nearest 15 neighbors

Table: Data Field Description - Customer Churn Dataset

	Variable Name	Description	Comments
1	CLIENTNUM	Client number. Unique identifier for the customer holding one credit card	Exclude from analysis
2	Attrition_Flag	This is the class variable or the target variable. If the account is closed then 1 else 0	target variable - binary
3	Customer_Age	Customer's age in years.	feature, numeric - continuous
4	Gender	Customer's gender	feature, categorical - nominal
5	Dependent_Count	Number of dependents	feature, numeric - discrete
6	Education_Level	Educational qualification of the account holder. For example, high school, college graduate, etc	feature, categorical - ordinal
7	Marital_Status	Married, Single, Divorced, Unknown	feature, categorical - nominal
8	Income_Category	Annual income of the account holder	feature, categorical - ordinal
9	Card_Category	Type of the card that the customer holds or has held	feature, categorical - ordinal
10	Months_on_book	Period of relationship with the bank	feature, numeric - continuous
11	Total_Relationship_Count	Total number of products of the bank held by the customer	feature, numeric - discrete
12	Months_Inactive_12_mon	Number of months inactive in the last 12 months	feature, numeric - continuous
13	Contacts_Count_12_mon	Number of contacts in the last 12 months	feature, numeric - discrete
14	Credit_Limit	Credit limit on the credit card	feature, numeric - continuous
15	Total_Revolving_Bal	Total revolving balance on the credit card	feature, numeric - continuous
16	Avg_Open_To_Buy	The average of open to buy credit line in the last 12 months	feature, numeric - continuous
17	Total_Amt_Chng_Q4_Q1	Change in transaction amount from Q4 to Q1	feature, numeric - continuous

18	Total_Trans_Amt	Total transaction amount in the last 12 months	feature, numeric - continuous
19	Total_Trans_Ct	Total transaction count in the last 12 months	feature, numeric - continuous
20	Total_Ct_Chng_Q4_Q1	Change in transaction count Q4 over Q1	feature, numeric - continuous
21	Avg_Utilization_Ratio	Average card utilization ratio over the last 12 months	feature, numeric - continuous