A Comprehensive guide to Analytics & Exploratory analysis

About Me


Planning
Operations
Revenue Management
Management
Econometrics
Hospitality Industry
Strategy
Sales & Marketing
MBA

Agenda

- A conceptual Frame work DCOVA
- Identifying the different types of data
- Selecting appropriate descriptive statistics (EDA)
- Choosing appropriate plot types (Visual Exploratory Data Analysis)
- Interpreting each plot type and inferring results

A Framework (5+2)

Define
Collect
Organize
Visualize
Analyze
+
Story telling with insights with applications
Iterate

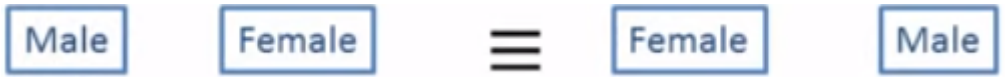
Data Types

- Categorical (Discrete)
  - Nominal
  - Ordinal
- Continuous
  - Interval
  - Ratio

Categorical - Nominal

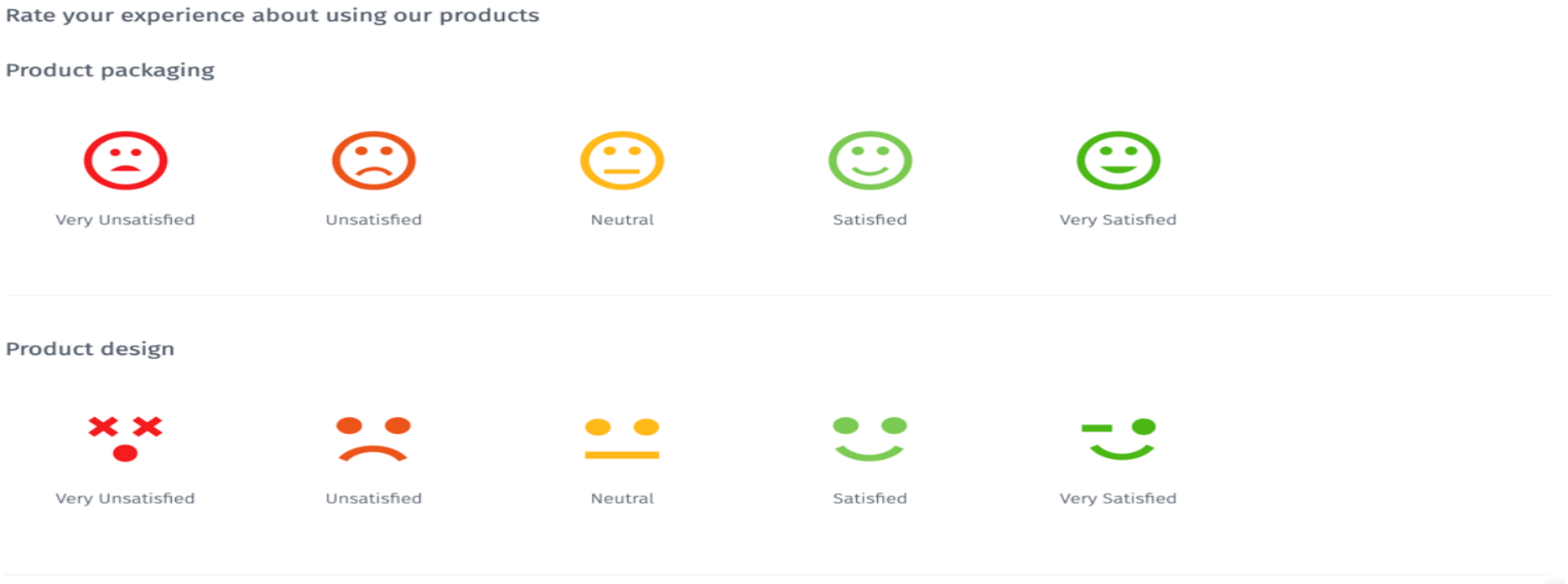
- Discrete data fields (Could be numeric or Alpha)
- No specific order to it. Can interchange them if wanted

- For example : Male could be taken first followed by Female or Vice versa



Categorical - Ordinal

- Discrete data fields
- Contains inherent order within it.



Continuous - Intervals

- They are numeric fields
- Distance between units are usually the same
- There is no value for Zero
  - Temperature of 0 deg does not mean there is no temperature
  - A person with an IQ score of 50 is not some one who is twice as smart someone with a score of 25

Continuous - Ratios

- They are Numeric fields
- Distance between units are the same
- Zero is considered Zero (Absolute Zero)
- Origin is Zero
  - Example :Zero sales means no Sales
  - 12 feet is twice as long as 6 feet

Lets identify the different types of Data

Country	Gender	Grade	Books	Math Score
Costa Rica	Male	9	0-10	410.66
Costa Rica	Female	9	0-10	343.99
Costa Rica	Male	10	11-25	418.92
Japan	Female	10	101-200	371.17
Turkey	Male	9	26-100	433.02
United States	Female	9	11-25	406.54
United States	Male	10	0-10	413.78

Descriptive Statistics

- Measures of Central Tendency
  - Mean
  - Median
  - Mode

- Measures of Variability
  - Range (Max – Min)
  - Standard Deviation $\sqrt{\frac{\sum_{i=1}^{N}(x_i-\overline{x})^2}{N-1}}$

Nominal Data

- Frequencies
  - Counting the number of events for each value
- Proportion
  - Dividing frequency of an item by the total frequency of all items
- Percentage
  - Multiplying Proportion times 100
- Visualize the above data using Bar charts

Ordinal Data

- Summarize
  - Frequencies
  - Proportion
  - Percentages
  - Mode
  - Median

- Visualize the above data using Bar charts

Continuous Data

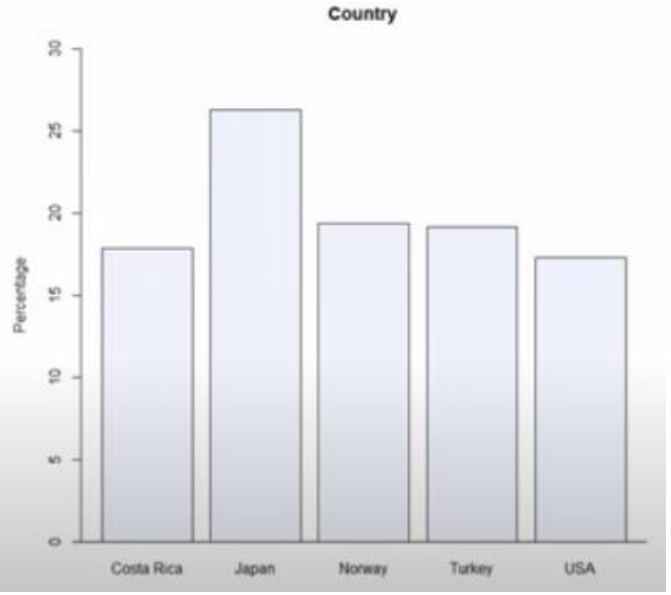
- Summarize
  - Mean
  - Median
  - Mode
  - Standard Deviation
  - Variance
- Visualize using
  - Box plots
  - Histograms

Categorical Summary

Table 1

Country of examinee

Country	Frequency	Percentage
Costa Rica	4,314	17.87
Japan	6,351	26.30
Norway	4,684	19.40
Turkey	4,618	19.13
USA	4,177	17.30



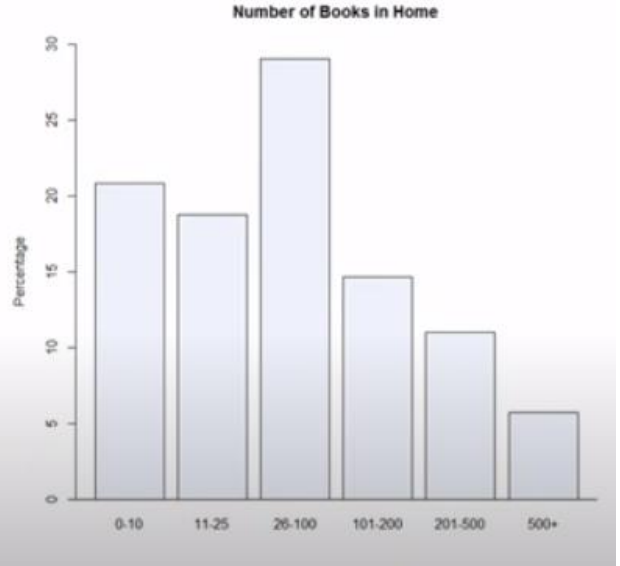
Order doesn't matter

Order matters

Table 2

Number of books in home

Books	Frequency	Percentage
0-10	4,901	20.85
11-25	4,411	18.77
26-100	6,828	29.05
101-200	3,440	14.64
201-500	2,580	10.98
500+	1,344	5.72



Relationship between two variables

- How does not variable impact another variable
- The types of variables matter
  - One Categorical Variable
    - Count of variable by field type
  - One Numeric Variable
    - Histogram / Box Plot
  - One categorical and one numeric variable
    - Pivot table summary / Side by side Box plots
  - 2 Categorical Variables
    - Pivot tables by counts
  - Two numeric Variables
    - Scatter plot

Outliers

- Need to ask Qs as to why outliers are occurring
- Do we keep them or remove them?
- What can outliers do to the data?
  - Scenario 1: 7,4,6,5,6,5,3,3,9,8
    - Mean = 5.6
    - Median = 5.5
    - Standard deviation = 2.01
  - Scenario 2: 7,4,6,5,6,5,3,3,20,8
    - Mean =6.7 (was 5.6)
    - Median = 5.5 (was 5.5)
    - Standard deviation = 4.94 (was 2.01)