EDA: Exploratory Data Analysis

EDA + DPP

Data Preprocessing

EDA Checklist

- Table-level analysis
- Per-column analysis and processing
- Cross-column analysis and processing
- Cross-table analysis

EDA: Table-Level Analysis

`df.describe()`

```
1 df = pd.read_csv('chocolate-bar-ratings.zip')
2 df.describe(include='all')
```

	Company \n(Maker-if known)	Specific Bean Origin\nor Bar Name	REF	Review\nDate	Cocoa\nPercent	Company\nLocation	Rating	Bean\nType	Broad Bean\nOrigin
count	1795	1795	1795.000000	1795.000000	1795	1795	1795.000000	1794	1794
unique	416	1039	NaN	NaN	45	60	NaN	41	100
top	Soma	Madagascar	NaN	NaN	70%	U.S.A.	NaN		Venezuela
freq	47	57	NaN	NaN	672	764	NaN	887	214
mean	NaN	NaN	1035.904735	2012.325348	NaN	NaN	3.185933	NaN	NaN
std	NaN	NaN	552.886365	2.927210	NaN	NaN	0.478062	NaN	NaN
min	NaN	NaN	5.000000	2006.000000	NaN	NaN	1.000000	NaN	NaN
25%	NaN	NaN	576.000000	2010.000000	NaN	NaN	2.875000	NaN	NaN
50%	NaN	NaN	1069.000000	2013.000000	NaN	NaN	3.250000	NaN	NaN
75%	NaN	NaN	1502.000000	2015.000000	NaN	NaN	3.500000	NaN	NaN
max	NaN	NaN	1952.000000	2017.000000	NaN	NaN	5.000000	NaN	NaN

EDA: Table-Level Analysis

```
df.head()      # First few rows of data

df.tail()      # Last few rows of data

df.sample(10) # Random sample of data

df.columns     # Column labels

df.dtypes      # Data types of columns
```

EDA: Per Column Analysis

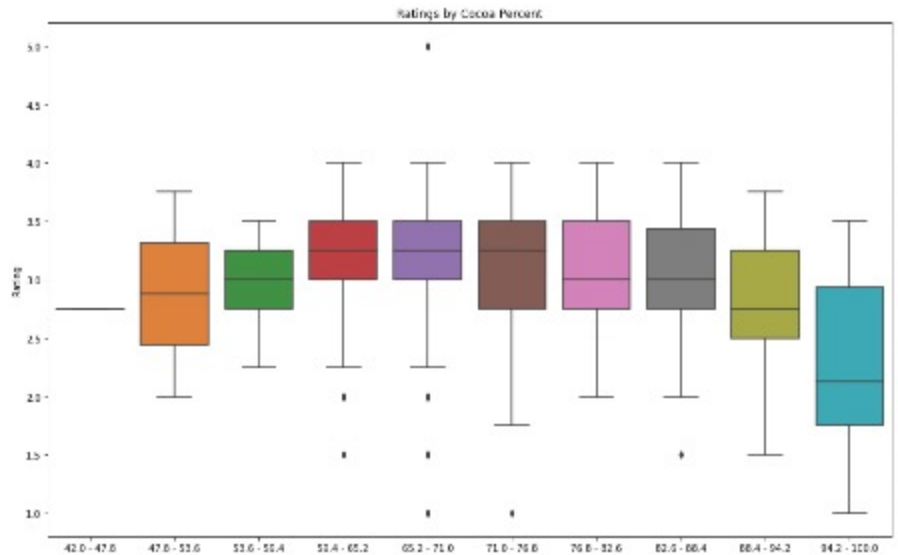
```
df[column].isna().any() # Missing values?

df[column].describe()   # Statistics for column

df[column].hist()       # Histogram

df[column].sample(10)   # Data sample
```

EDA: Cross-Column Analysis

```
1 df.corr()
```

	Cocoa Percent	REF	Review Date	Rating
Cocoa Percent	1.000000	0.037791	0.038177	-0.164820
REF	0.037791	1.000000	0.985076	0.101030
Review Date	0.038177	0.985076	1.000000	0.099846
Rating	-0.164820	0.101030	0.099846	1.000000

EDA: Cross-Column Analysis



EDA: Cross-Table Analysis

	Country	Latitude	Longitude
count	58	58.000000	58.00000
unique	58	NaN	NaN
top	South Korea	NaN	NaN
freq	1	NaN	NaN
mean	NaN	23.392586	-0.45931
std	NaN	29.271631	71.95592
min	NaN	-41.190000	-99.10000
25%	NaN	5.387500	-64.97250
50%	NaN	23.785000	1.07000
75%	NaN	50.395000	20.51250
max	NaN	64.100000	178.30000

EDA: Data Preprocessing

- Missing values
- Invisible values
- Placeholder values
- Values out of the expected range
- Incorrect data types
- Inconsistent values (USA vs. U.S.A.)

EDA: *Not Necessarily* Data Preprocessing

- Normalization
- Standardization
- One-hot encoding

Preprocessing Strategies

- Fill in missing data
- Fix broken data
- Remove problematic columns
- Remove problematic rows

Your Turn!

Chocolate Bar Ratings

- Lab Part 1: Individual tables and columns
- Lab Part 2: Cross tables and columns