Azercell

Exploratory Data Analysis

Azercell DataMinds Bootcamp

2025

Content

- Introduction
  - What is data
  - Why EDA is important

- Data Types
  - Data types and 4 levels of data
  - The Nominal level
  - The Ordinal level
  - The Interval level
  - The Ratio Level

- Data Analysis
  - Univariate data analysis
    - Introduction
    - Central tendency
    - Dispersion
    - Distribution & Skewness
    - Category features analysis
    - Outliers Detection

  - Multivariate Analysis
    - Correlation & Covariance
    - Pairplots & Scatterplots
    - Category vs Numerical
    - Grouped Aggregations
    - Category VS Category

Introduction

What is Data

The data representation

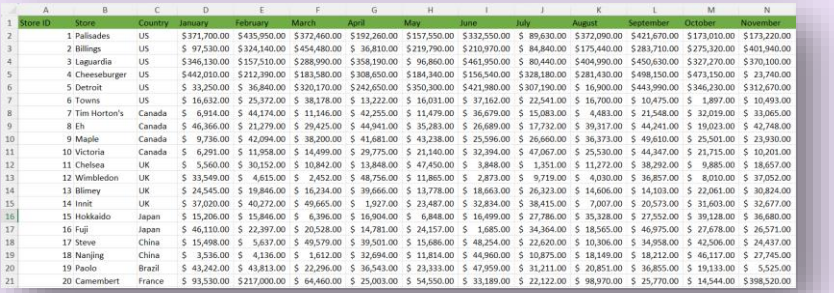
Data are a collection of discrete or continuous values that convey information, describing the quantity, quality, fact, statistics, other basic units of meaning, or simply sequences of symbols that may be further interpreted formally. *(from Wikipedia)*

Structured data	Unstructured data
Tabular data can be represented in rows and columns and in a relational database.	Cannot be represented as tabular data. Generally thought of as a mass of data.
Generally easier to use as impute data for machine learning.	Harder to use as impute data, as they need to be structured first.
Require less storage capacity.	Require more storage capacity.
Estimated 20% of data in the enterprise.	Estimated 80% of data in the enterprise.

- **Data** = Facts, figures, and information collected for reference or analysis.
- Two main types
  - — Structured: tables (CSV, Excel, SQL)
  - — Unstructured: Text images videos
- Common Formats: CSV, JSON, Excel, SQL, Parquet
- Dimensions: Rows = observation, Column – features/variables

Unstructured

Structured



	Store ID	Store	Country	January	February	March	April	May	June	July	August	September	October	November
2	1	Palisades	US	$371,700.00	$435,950.00	$372,460.00	$192,260.00	$157,550.00	$332,550.00	$ 89,630.00	$372,090.00	$421,670.00	$173,010.00	$173,220.00
3	2	Billings	US	$ 97,530.00	$324,140.00	$454,480.00	$ 36,810.00	$219,790.00	$210,970.00	$ 84,840.00	$175,440.00	$283,710.00	$275,320.00	$401,940.00
4	3	Laguardia	US	$346,130.00	$157,510.00	$288,990.00	$358,190.00	$ 96,860.00	$461,950.00	$ 80,440.00	$404,990.00	$450,630.00	$327,270.00	$370,100.00
5	4	Cheeseburger	US	$442,010.00	$212,390.00	$183,580.00	$308,650.00	$184,340.00	$156,540.00	$328,180.00	$281,430.00	$498,150.00	$473,150.00	$ 23,740.00
6	5	Detroit	US	$ 33,250.00	$ 36,840.00	$320,170.00	$242,650.00	$350,300.00	$421,980.00	$307,190.00	$ 16,900.00	$443,990.00	$346,230.00	$312,670.00
7	6	Towns	US	$ 16,632.00	$ 25,372.00	$ 38,178.00	$ 13,222.00	$ 16,031.00	$ 37,162.00	$ 22,541.00	$ 16,700.00	$ 10,475.00	$ 1,897.00	$ 10,493.00
8	7	Tim Horton's	Canada	$ 6,914.00	$ 44,174.00	$ 11,146.00	$ 42,255.00	$ 11,479.00	$ 36,679.00	$ 15,083.00	$ 4,483.00	$ 21,548.00	$ 32,019.00	$ 33,065.00
9	8	Eh	Canada	$ 46,366.00	$ 21,279.00	$ 29,425.00	$ 44,941.00	$ 35,283.00	$ 26,689.00	$ 17,732.00	$ 39,317.00	$ 44,241.00	$ 19,023.00	$ 42,748.00
10	9	Maple	Canada	$ 9,736.00	$ 42,094.00	$ 38,200.00	$ 41,681.00	$ 43,238.00	$ 25,596.00	$ 26,660.00	$ 36,373.00	$ 49,610.00	$ 25,501.00	$ 23,930.00
11	10	Victoria	Canada	$ 6,291.00	$ 11,958.00	$ 14,499.00	$ 29,775.00	$ 21,140.00	$ 32,394.00	$ 47,067.00	$ 25,530.00	$ 44,347.00	$ 21,715.00	$ 10,201.00
12	11	Chelsea	UK	$ 5,560.00	$ 30,152.00	$ 10,842.00	$ 13,848.00	$ 47,450.00	$ 3,848.00	$ 1,351.00	$ 11,272.00	$ 38,292.00	$ 9,885.00	$ 18,657.00
13	12	Wimbledon	UK	$ 33,549.00	$ 4,615.00	$ 2,452.00	$ 48,756.00	$ 11,865.00	$ 2,873.00	$ 9,719.00	$ 4,030.00	$ 36,857.00	$ 8,010.00	$ 37,052.00
14	13	Blimey	UK	$ 24,545.00	$ 19,846.00	$ 16,234.00	$ 39,666.00	$ 13,778.00	$ 18,663.00	$ 26,323.00	$ 14,606.00	$ 14,103.00	$ 22,061.00	$ 30,824.00
15	14	Innit	UK	$ 37,020.00	$ 40,272.00	$ 49,665.00	$ 1,927.00	$ 23,487.00	$ 32,834.00	$ 38,415.00	$ 7,007.00	$ 20,573.00	$ 31,603.00	$ 32,677.00
16	15	Hokkaido	Japan	$ 15,206.00	$ 15,846.00	$ 6,396.00	$ 16,904.00	$ 6,848.00	$ 16,499.00	$ 27,786.00	$ 35,328.00	$ 27,552.00	$ 39,128.00	$ 36,680.00
17	16	Fuji	Japan	$ 46,110.00	$ 22,397.00	$ 20,528.00	$ 14,781.00	$ 24,157.00	$ 1,685.00	$ 34,364.00	$ 18,565.00	$ 46,975.00	$ 27,678.00	$ 26,571.00
18	17	Steve	China	$ 15,498.00	$ 5,637.00	$ 49,579.00	$ 39,501.00	$ 15,686.00	$ 48,254.00	$ 22,620.00	$ 10,306.00	$ 34,958.00	$ 42,506.00	$ 24,437.00
19	18	Nanjing	China	$ 3,536.00	$ 4,136.00	$ 1,612.00	$ 32,694.00	$ 11,814.00	$ 44,960.00	$ 10,875.00	$ 18,149.00	$ 18,212.00	$ 46,117.00	$ 27,745.00
20	19	Paolo	Brazil	$ 43,242.00	$ 43,813.00	$ 22,296.00	$ 36,543.00	$ 23,333.00	$ 47,959.00	$ 31,211.00	$ 20,851.00	$ 36,855.00	$ 19,133.00	$ 5,525.00
21	20	Camembert	France	$ 93,530.00	$217,000.00	$ 64,460.00	$ 25,003.00	$ 54,550.00	$ 33,189.00	$ 22,122.00	$ 98,970.00	$ 25,770.00	$ 14,544.00	$398,520.00

Azercell

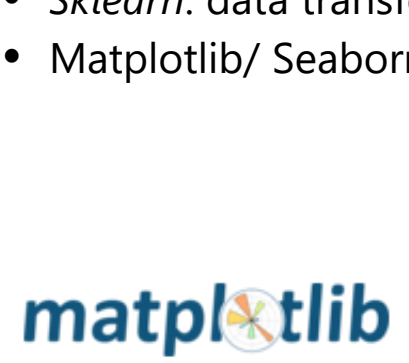
Why is EDA important

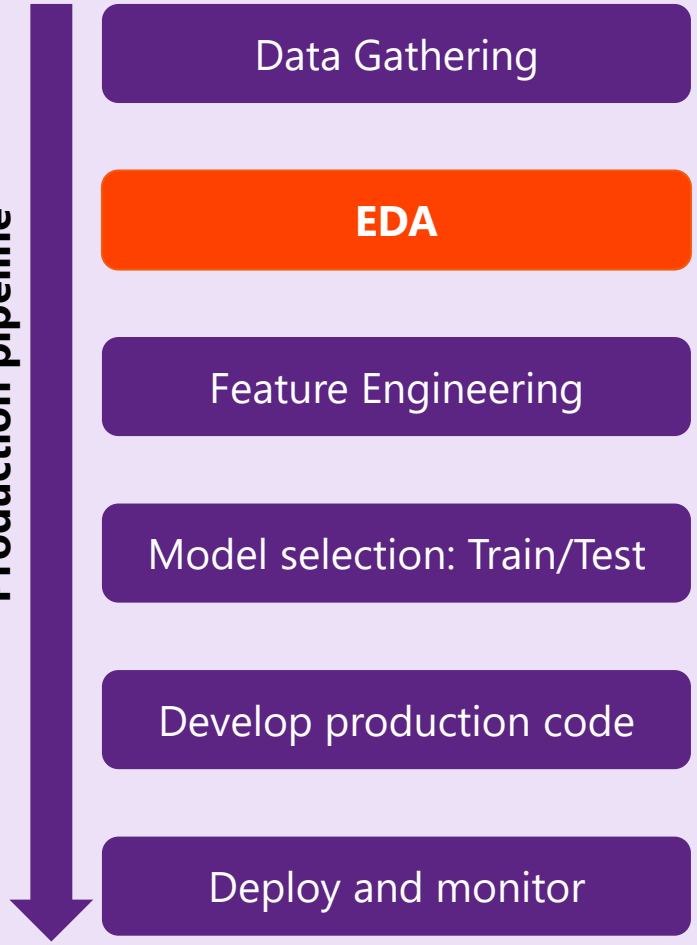
Key important features

- Understand the **structure** and **patterns** in the data.
- Identify **errors**, **outliers**, and **missing values**.
- Determine **relationships** between variables.
- Guide **feature selection** and **modeling decisions**.
- Example: Avoiding garbage-in, garbage-out in machine learning.

Tools

- *Pandas*: for structured data (row/column) conception
- *Numpy*: statistics
- *Sklearn*: data transformation/ imputation
- Matplotlib/ Seaborn: visualization



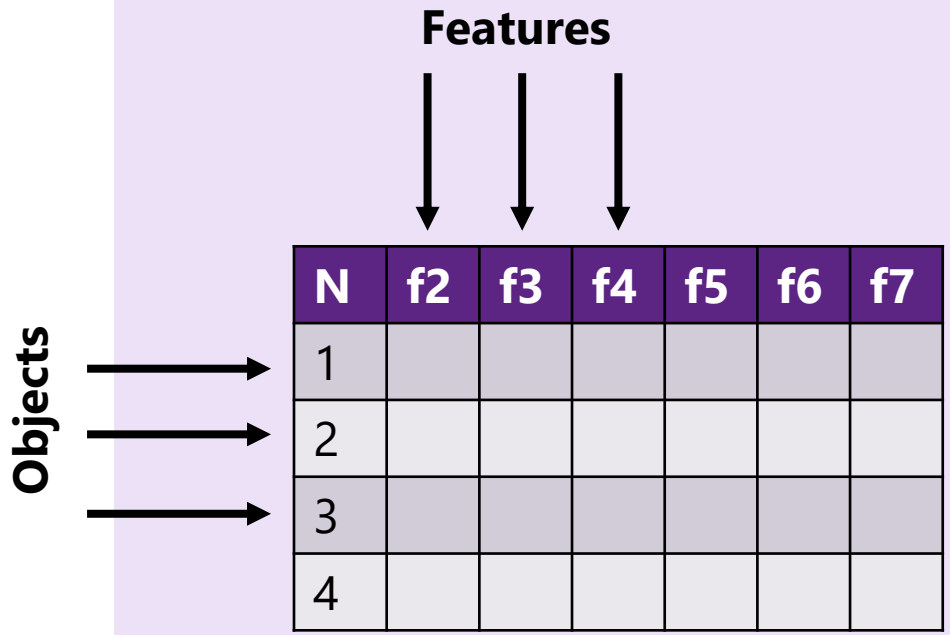


Azercell

Data Types

Data Types

Feature can be either quantitative (i.e., numerical in nature) or qualitative (i.e., categorical in nature)



Four levels of Data

Knowing what level your data belong to makes all the difference in deciding what kind of feature engineering technique is possible, let alone appropriate.

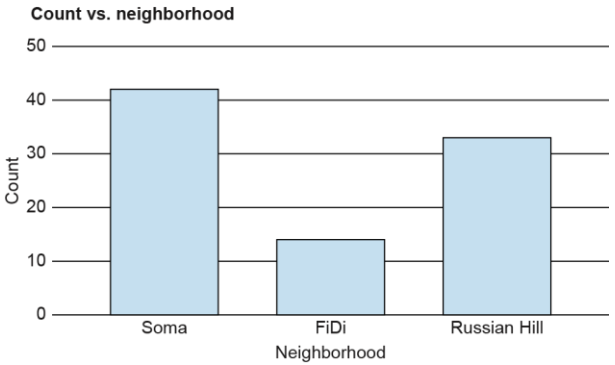
- The Nominal level
- The Ordinal level
- The Interval level
- The Ratio level

Data can exist on exactly one of these four levels, and knowing which level we are working with for each of our features will often dictate what kinds of operations are and aren't allowed to be used on them

The Nominal level

Data at the nominal level are qualitative through and through. This includes categories, labels, descriptions, and classifications of things that involve no quantitative meaning whatsoever and have no discernable order.

Type of home	Neighborhood
Apartment	Soma
Single family	FiDi
Apartment	Russian Hill
Condo	Soma
Single family	Soma
Single family	FiDi
Apartment	Soma
Duplex	Russian Hill
Apartment	Soma
Apartment	FiDi
Condo	FiDi
Apartment	Russian Hill
Apartment	Soma
...	....



Dealing with Nominal data

Binary Encoding

Types of connetion
4G
3G
3G
5G
4G

3G	4G	5G
0	1	0
1	0	0
1	0	0
0	0	1
0	1	0

More complicated forms of transforming Nominal data will be presented *Feature engineering*.

There're not so much math operations

Example:

- City
- Blood types (you cannot take average)
- Region
- Phone brands
- Color

```
>>> s1 = ['a', 'b', np.nan]
>>> pd.get_dummies(s1)
a b
0 True False
1 False True
2 False False
```

Azercell

The Ordinal level

Ordinal data depict qualitative data with some sense of order but stop short of having meaningful differences between values. One of the most common examples of data on the ordinal level is customer support satisfaction surveys.

still a category, but we have a sense of order

Example:

- Very unhappy
- Unhappy
- Neutral
- Happy
- Very Happy

don't have an easy way of defining the space between happy and very happy

Dealing with Ordinal level data

Assigning some weights:

```
import pandas as pd

df = pd.DataFrame(
        {
        'Condition': ["Happy", "VeryHappy", "Unhappy", "Unhappy", "Happy"],
        'UserID': [1, 2, 3, 4, 5]
        }
)

df["Condition"] = pd.to_numeric(df['Condition'].replace({
        "VeryHappy": "2",
        "Happy": "1",
        "Unhappy": "-1",
        "Neutral": "0",
        "VeryUnhappy":"-2",
}))
```

Azercell

Interval level

interval level are similar to data at the ordinal level, except for the crucial fact that differences between values have a consistent meaning.

IMPORTANT!

data at the interval level **do NOT have** is the concept of a true zero.

the concept of 0 degrees does not indicate "an absence of temperature"

Dealing with interval level data

When we have the ability to add and subtract numbers and can rely on those additions or subtractions being consistent, we can start to calculate things like arithmetic *means, medians*, and *standard deviations*.

Example:

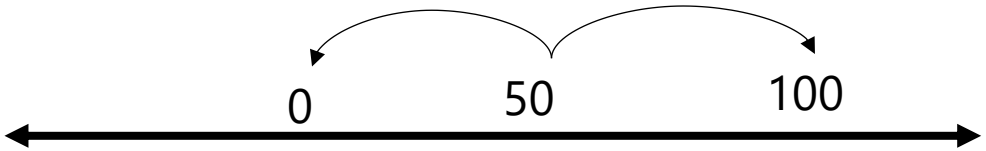
temperature. We clearly have a sense of order—68 degrees is hotter than 58 degrees

Survey from the ordinal level at the interval level if we choose to give meaning to differences between survey results

The Ratio level

This is the scale of data most people think of when they think of qualitative data. Data at the ratio level are, as you've probably already guessed, identical to data at the interval level with a **true zero existing.**

At the interval level, we can only add and subtract values together with meaning. At the ratio level, with the concept of a **true zero**, we can divide and multiply values together and have their results be meaningful.



Twice as much as 50
Means twice as far from
0 than 50 is

Type of mean	Description	Level of data	When to use	When not to use
Arithmetic mean	Additive mean	Interval and ratio	•Data have consistent units. •Data are additive in nature.	•When we don't want our mean to be skewed by outliers.
Geometric mean	Multiplicative mean	Ratio	•Data are multiplicative in nature. •Data are on different scales or have differing units.	•If scales and units matter, the geometric mean will obscure them. •Data have zeros or negative values.
Harmonic mean	The reciprocal of the arithmetic mean of the reciprocals of the data	Ratio	•Data values are ratios (fractions of other values).	•Difficult to interpret for those who are unfamiliar with the harmonic mean. •Data have zeros or negative values.

Example:

Money—We can define a true zero as being the absence of money. We don't have any money if we have 0 dollars or 0 AZN.

Age, *height*, and *weight*—These would also count as being on the ratio level.

One hundred dollars is twice as much as $50, and $250 is half as much as $500. These sentences have meaning because we can visualize the concept of having no money

Data Analysis

Introduction to Univariate analysis

Univariate analysis focuses on analyzing a single variable. It provides a foundation for understanding the nature of the data, such as its distribution, central tendency, and variability. This step is essential before moving to complex multivariate interactions.

Objectives:
•Identify data type (numerical/categorical)
•Understand basic structure of data
•Spot missing values or outliers

Key Measures:
•Central Tendency: Mean, Median, Mode
•Dispersion: Variance, Standard Deviation, Range

```
df['Age'].describe()
```

```
count 714.000000 mean
29.699118 std
14.526497 min
0.420000
25% 20.125000 50%
28.000000 75%
38.000000 max
80.000000 Name: Age,
dtype: float64
```

Azercell

Univariate Analysis: Central tendency

Central tendency represents the "center" of a dataset. The three primary measures—mean, median, and mode—each describe this center differently. The mean is the arithmetic average, the median is the midpoint, and the mode is the most frequently occurring value.

Understanding the central tendency helps in grasping how the data behaves. For example, if your mean and median are very different, the data may be skewed. In real-world applications like pricing or income, median is often more robust.

Formulas:
•Mean:
Median: Middle value of sorted data
•Mode: Most frequent value
Why Important?
•Mean sensitive to outliers
•Median gives robust center
•Mode useful for categorical data

```
print("Mean:", df['Fare'].mean())
print("Median:", df['Fare'].median())
print("Mode:", df['Fare'].mode()[0])
```

Azercell

Univariate Analysis: Dispersion (Spread)

Dispersion tells us how spread out the data is around the central value. Variance and standard deviation are the two main measures that quantify this spread mathematically. A low standard deviation means the data is clustered close to the mean.

Range and Interquartile Range (IQR) are also useful metrics. While range gives a simple max-min difference, IQR shows where the central 50% of values lie, giving insight into data consistency and outlier existence.

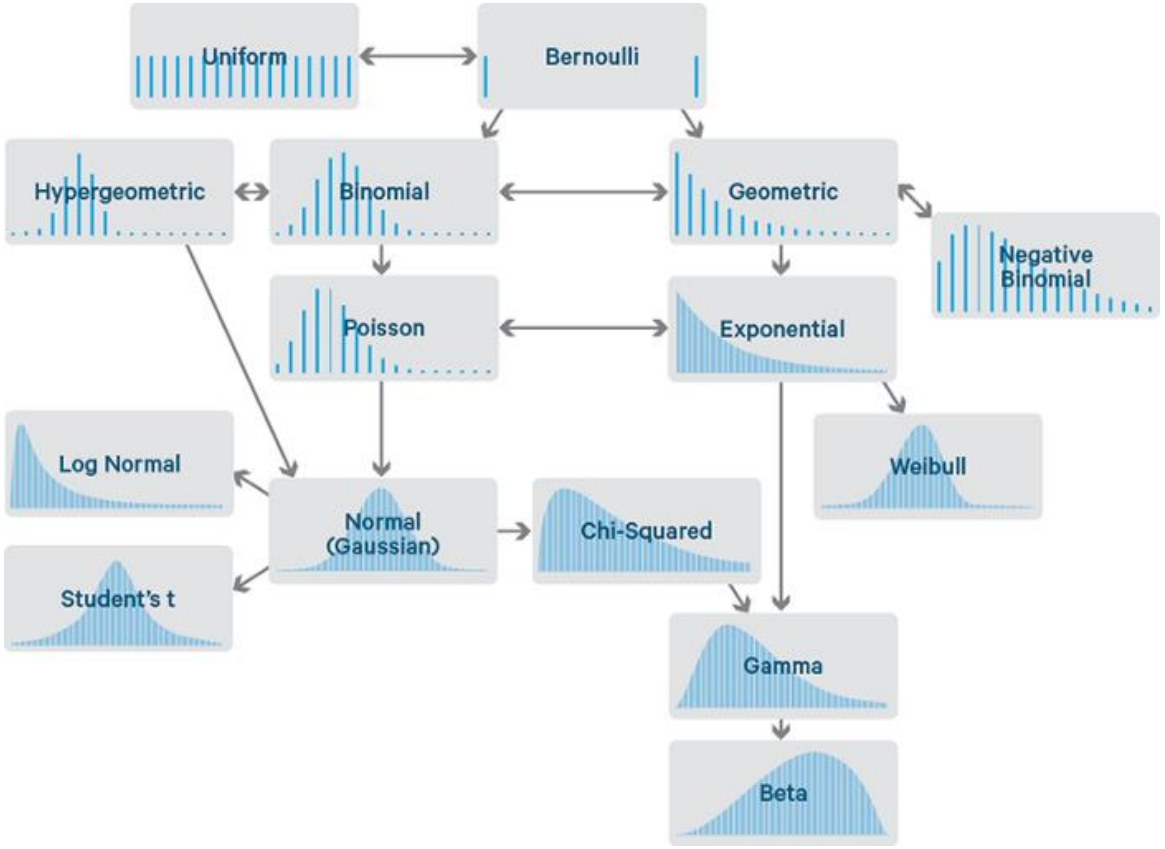
High dispersion may indicate variability in customer behavior, environmental conditions, or product performance.
Understanding spread is crucial for risk analysis and anomaly detection.

```
print("STD:", df['Fare'].std())
print("Variance:", df['Fare'].var())
```

Azercell

Univariate Analysis: Distribution & Skewness

The distribution of a variable tells you how its values are spread across possible outcomes. Histograms and Kernel Density Estimation (KDE) plots are useful to visualize this. A normal (bell-shaped) distribution is often a baseline assumption in statistical modeling.



Skewness quantifies the asymmetry of the distribution. Right-skewed data (positive skew) means a long tail on the right—common in income data. Left-skewed distributions are less frequent but also important.

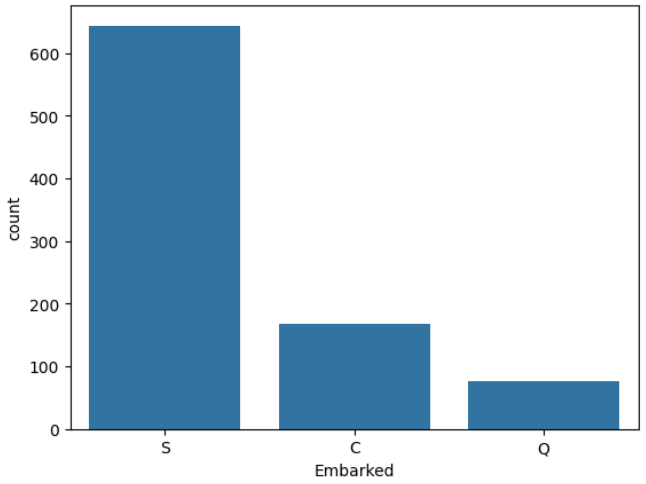
Azercell

Univariate Analysis: Categorical features

Categorical variables like gender or city codes need a different treatment from numeric data. Here, we're more interested in frequency and proportions than averages. Mode becomes the most meaningful central tendency.

Understanding category distribution can guide strategies like data balancing or feature encoding in machine learning. Always explore these before jumping into modeling.

```
sns.countplot(x='Embarked', data=df)
```



Azercell

Univariate Analysis: Outlier Detection

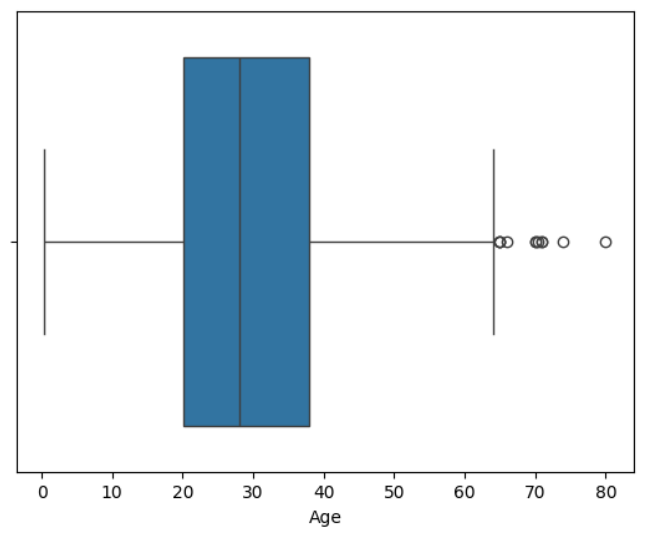
Outliers are values that are significantly different from the rest of the data. These could be legitimate values (e.g., a billionaire in income data) or errors. They can distort means and impact model performance.

The boxplot visualizes outliers and IQR-based thresholds. Anything below Q1 − 1.5×IQR or above Q3 + 1.5×IQR is considered an outlier. This method is robust and works well for most datasets.

```
Q1 = df['Fare'].quantile(0.25)
Q3 = df['Fare'].quantile(0.75)
IQR = Q3 - Q1
outliers = df[(df['Fare'] < Q1 - 1.5*IQR) | (df['Fare'] > Q3 + 1.5*IQR)]
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
27	28	0	1	Fortune, Mr. Charles Alexander	male	19.0	3	2	19950	263.0000	C23 C25 C27	S
31	32	1	1	Spencer, Mrs. William Augustus (Marie Eugenie)	female	NaN	1	0	PC 17569	146.5208	B78	C
34	35	0	1	Meyer, Mr. Edgar Joseph	male	28.0	1	0	PC 17604	82.1708	NaN	C
52	53	1	1	Harper, Mrs. Henry Sleeper (Myna Haxtun)	female	49.0	1	0	PC 17572	76.7292	D33	C

```
sns.boxplot(x=df['Age'])
```



Azercell

Multivariate Analysis: introduction

Multivariate analysis looks at how multiple variables relate to each other. This is where EDA moves from simple summarization to finding patterns and interactions.
This kind of analysis can reveal dependencies, correlations, and potential causal relationships. It's essential in tasks like feature engineering and dimensionality reduction.

Study relationships between **two or more features**.

- Num vs Num
- Cat vs Num
- Cat vs Cat

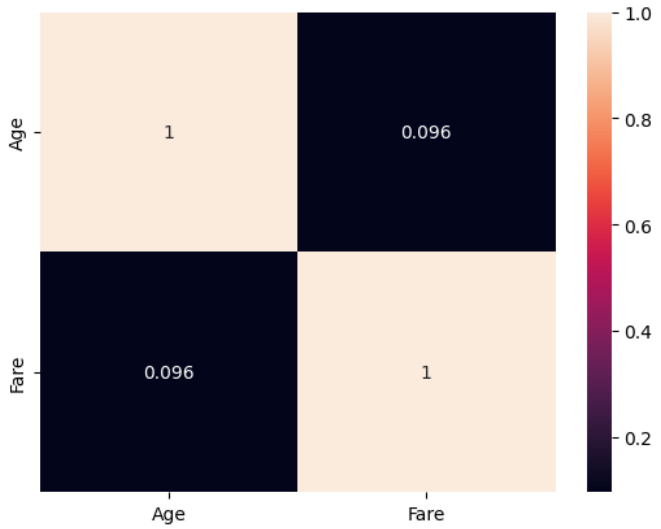
Common Questions:
•Does one feature influence another?
•Are features correlated?

Multivariate Analysis: Correlation & Covariance

Correlation and covariance measure how variables move together. Covariance shows direction, while correlation (normalized) shows both direction and strength, with values from -1 to +1.

A high positive correlation (close to +1) means variables increase together. A negative correlation (close to -1) means one increases while the other decreases. Zero means no linear relationship.

```
sns.heatmap(df[['Age', 'Fare']].corr(), annot=True)
```




```
pd.crosstab(
  df['Pclass'],
  df['Survived'],
  normalize='index'
)
```

Survived	0	1
Pclass		
1	0.370370	0.629630
2	0.527174	0.472826
3	0.757637	0.242363

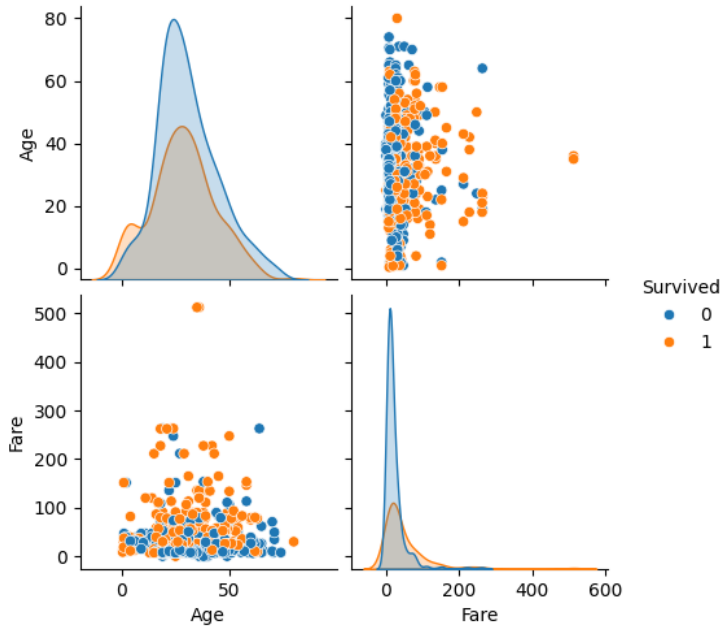
Azercell

Multivariant Analysis: Pairplots & Scatterplots

Pairplots provide a visual overview of the relationships between multiple numeric variables. They are perfect for initial pattern detection and understanding class separation.

Scatter plots show relationships between two variables. When colored by a third (like a category), they provide insight into group-wise patterns.

```
sns.pairplot(df[['Age', 'Fare', 'Survived']], hue='Survived')
```

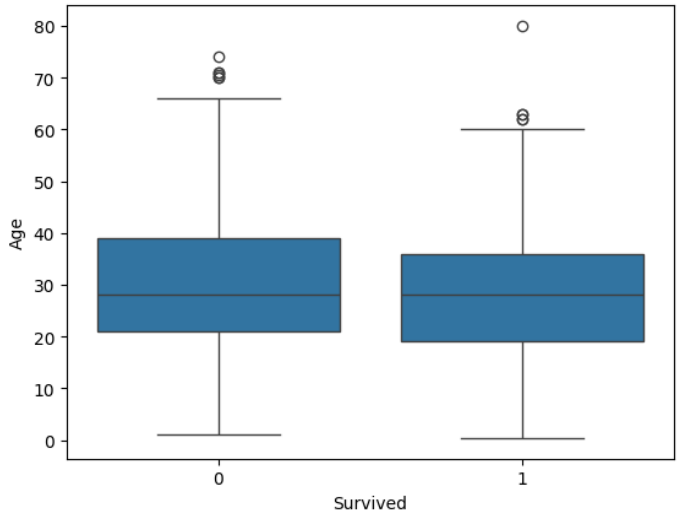


Multivariant Analysis: Categorical vs Numerical

When comparing a categorical variable (like class or gender) against a numerical one (like income or age), boxplots and violin plots are the go-to visual tools.

For example, if "Survived" affects the distribution of "Fare", it could mean Fare is an important predictor of survival in the Titanic dataset.

```
sns.boxplot(x='Survived', y='Age', data=df)
```



Azercell

Multivariant Analysis: Grouped Aggregations

groupby() operations summarize data at different aggregation levels. This is essential when you want to compare averages, counts, or medians across groups.

For example, comparing average fare across passenger classes helps identify if higher class passengers paid more. This can help with segment-wise strategies.

Why Use?

- Summarize behavior per group
- Compare central tendencies

```
df.groupby(
   ['Pclass','Survived']
)['Age'].median()
```

```
Pclass Survived
1 0 45.25
  1 35.00
2 0 30.50
  1 28.00
3 0 25.00
  1 22.00
Name: Age, dtype: float64
```

```
df.groupby('Pclass')['Fare'].mean()
```

```
Pclass
1 84.154687
2 20.662183
3 13.675550
Name: Fare, dtype: float64
```

Azercell

Multivariant Analysis: Categorical vs Categorical

Analyzing relationships between two categorical variables helps you uncover hidden group-level patterns. Crosstab is a handy tool for this. A heatmap version of the crosstab makes it easy to visually compare proportions or frequencies across combinations of categories.

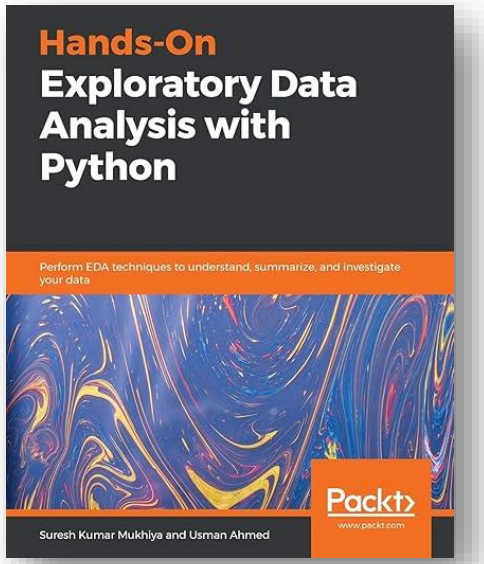
This kind of analysis is crucial in understanding class balance and for feature engineering, like combining categorical variables or encoding them more effectively.

```
pd.crosstab(df['Pclass'], df['Survived’], normalize='index')
```

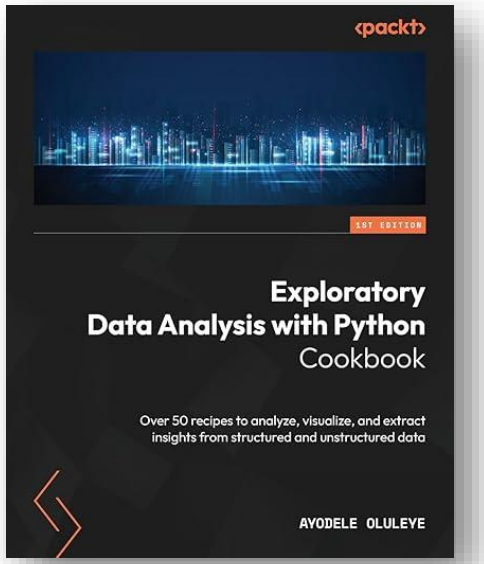
Survived	0	1
Pclass		
1	0.370370	0.629630
2	0.527174	0.472826
3	0.757637	0.242363

Azercell

Sources:



Hands-On Exploratory Data Analysis with Python
Suresh Kumar Mukhiya ,
Usman Ahmed



Exploratory Data Analysis with Python Cookbook
Ayodele Oluleye

Links

Data Visualization using Matplotlib
https://medium.com/data-science/data-visualization-using-matplotlib-16f1aae5ce70

Thank you for your attention