Data Analytics with Python

Lecture 3

Ilias Suvanov

ilias.suvanov@gmail.com

2020 年 10 月 21 日

Exploratory Data Analysis

Content

Basic Statistics

Max and Min

Max

Maximal value among numerical data points.

```
arr = np.array([1,4,3,5])
np.amax(arr)
```

Min

Minimal value among data numerical points.

```
arr = np.array([1,4,3,5])
np.amin(arr)
```

Mode

The most commonly occurring category or value in a dataset.

Mean

Mean (or Average)

The sum of all values divided by the number of values.

Forumula $Mean = \frac{\sum_{i=1}^{N} x_i}{N}$

```
ls = [171,175,174,180,165,163]
arr = np.array(ls)
np.mean(arr)
```

Weighted Mean

The sum of all values times a weight divided by the sum of the weights.

Formula: $WeightedMean = \bar{x}_w = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i}$

```
np.average(np.arange(1, 11), weights=np.arange(10, 0, -1))
```

Trimmed Mean (or Truncuted Mean)

The average of all values after dropping a fixed number of extreme values.

```
# Return mean of array after trimming distribution from both tails.
# If proportiontocut = 0.1, slices off 'leftmost' and 'rightmost' 10% of scores.
from scipy import stats
x = np.arange(10)
stats.trim_mean(x, 0.1)
```

Median

Median

The value such that one-half of the data(after being sorted) lies above and below.

```
1  arr = np.array([10, 7, 4, 8, 3, 1, 2])
2  np.median(arr)
```

Median some times more useful than mean.

Terminology

Robust - Not sensitive to extreme values.
Outlier-A data value that is very different from most of the data.

Summary

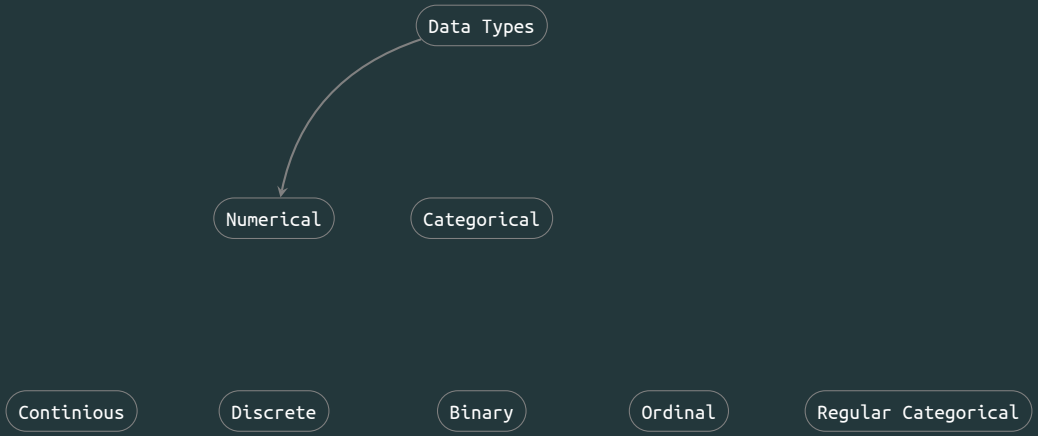
- The basic metric for location is the mean, but it can be sensitive to extreme val□ ues (outlier)

Summary

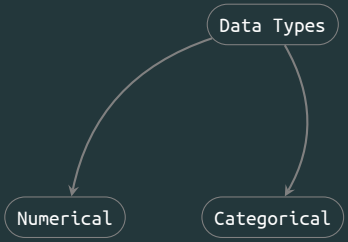
- The basic metric for location is the mean, but it can be sensitive to extreme val ues (outlier)
- Other metrics (median, trimmed mean) are more robust

Data Types

Data Types


Data Types
Numerical
Categorical
Continious
Discrete
Binary
Ordinal
Regular Categorical

Data Types



Continious

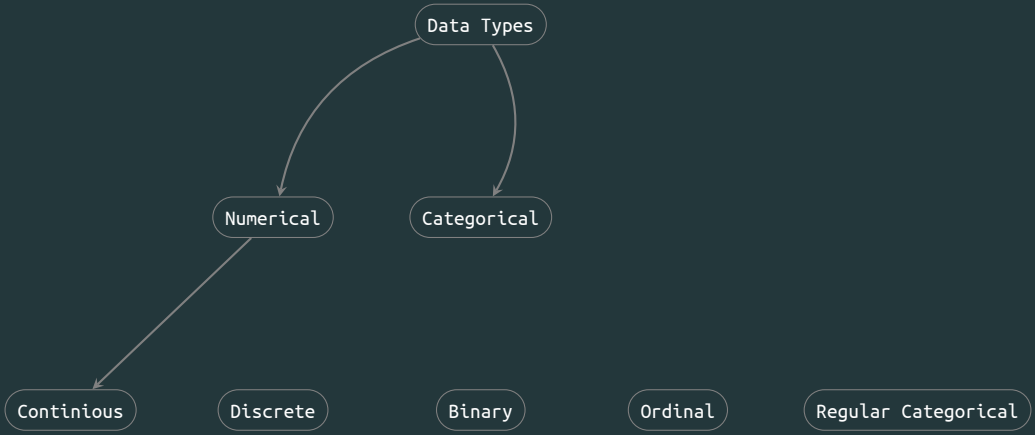
Discrete

Binary

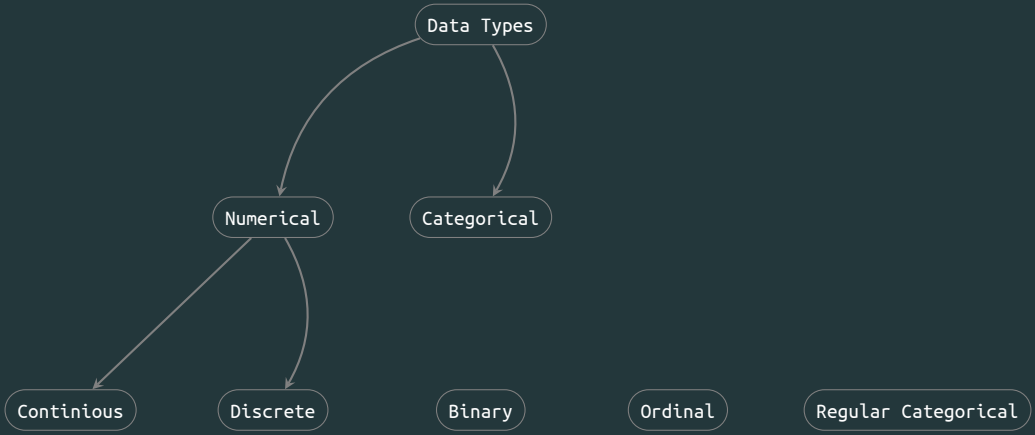
Ordinal

Regular Categorical

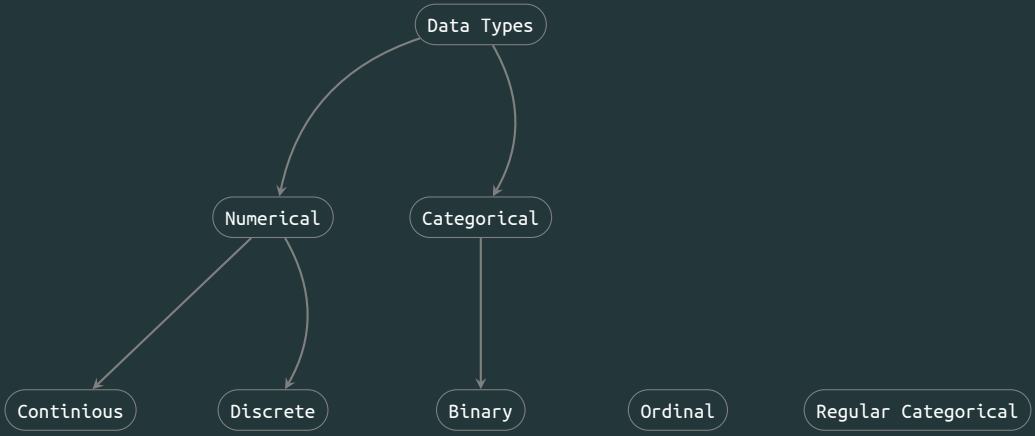
Data Types


Data Types
Numerical
Categorical
Continious
Discrete
Binary
Ordinal
Regular Categorical

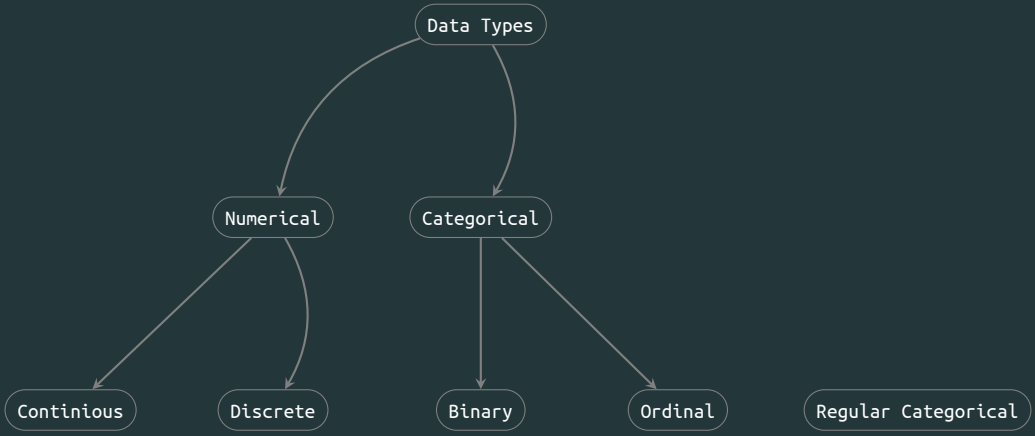
Data Types


Data Types
Numerical
Categorical
Continious
Discrete
Binary
Ordinal
Regular Categorical

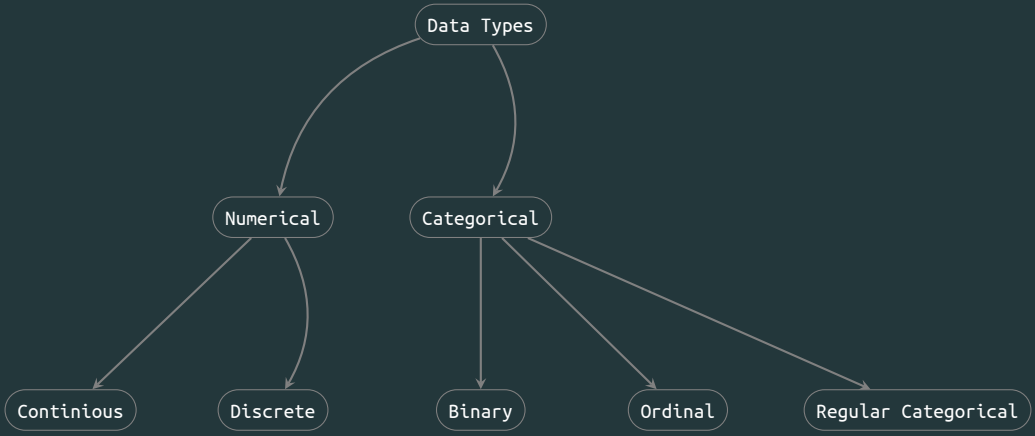
Data Types


Data Types
Numerical
Categorical
Continious
Discrete
Binary
Ordinal
Regular Categorical

Data Types


Data Types
Numerical
Categorical
Continious
Discrete
Binary
Ordinal
Regular Categorical

Data Types


Data Types
Numerical
Categorical
Continious
Discrete
Binary
Ordinal
Regular Categorical

Summary

- The data types is important to help determine the type of visual display.

Summary

- The data types is important to help determine the type of visual display.
- To let software know what statistical model it should apply for data analysis.

Variability

Estimates of Variability

Location (Mean, Median) is just one dimension in summarizing a feature. A second dimension, variability, also referred to as dispersion, measures whether the data values are tightly clustered or spread out. At the heart of statistics lies variability: measuring it, reducing it, distinguishing random from real variability, identifying the various sources of real variability and making decisions in the presence of it

Deviations

The difference between the observed values and the estimate of location.

Range

The difference between the largest and the smallest value in a data set

Variance

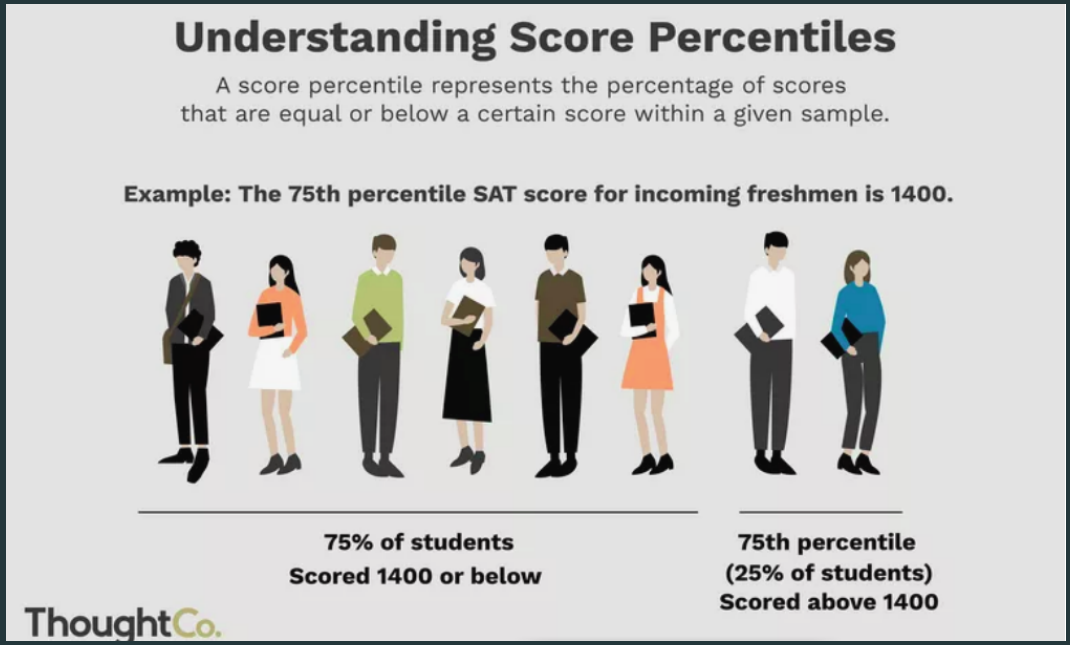
The sum of squared deviations from the mean divided by N-1 where N is the number of data values.

$$Variance = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}$$

Standard Deviation

The square root of the variance.

Percentiles



Percentiles

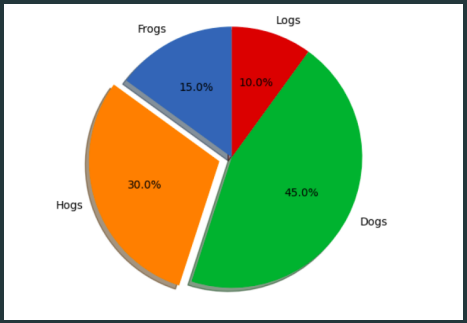
The P-th percentile is a value such that at least P per cent of the values take on this value or less and at least (100-P) percent of the values take on this value or more.

For example, to find the 80th percentile, sort the data. Then, starting with the smallest value, proceed 80 percent of the way to the largest value. Note that the median is the same thing as the 50th percentile.

Data Vizualization

Pie Chart

The frequency or proportion for each category plotted as wedges in a pie.

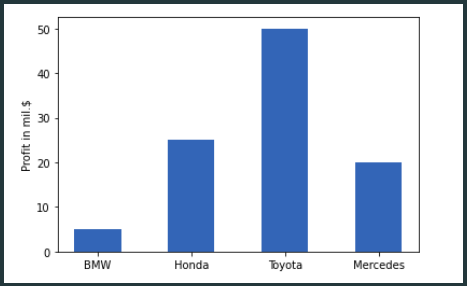


Code

```
1  import matplotlib.pyplot as plt
2  # Pie chart, where the slices will
3  # be ordered and plotted counter-clockwise:
4  labels = 'Frogs', 'Hogs', 'Dogs', 'Logs'
5  sizes = [15, 30, 45, 10]
6  explode = (0, 0.1, 0, 0)
7  # only "explode" the 2nd slice (i.e. 'Hogs')
8  fig1, ax1 = plt.subplots()
9  ax1.pie(sizes, explode=explode, labels=labels
10     shadow=True, startangle=90)
11 ax1.axis('equal')
12 # Equal aspect ratio ensures
13 # that pie is drawn as a circle.
14 plt.show()
```

Bar Chart

The frequency or proportion for each category plotted as bars.

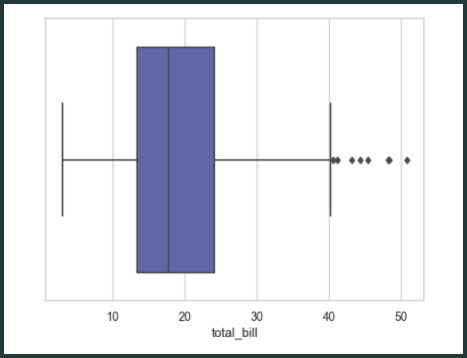


Code

```
1  import matplotlib.pyplot as plt
2  data = [5., 25., 50., 20.]
3  names = ['BMW', 'Honda', 'Toyota',\\ 'Mercede
4  plt.bar(names, data, width = 0.5)
5  plt.ylabel('Profit in mil.$')
6  plt.show()
```

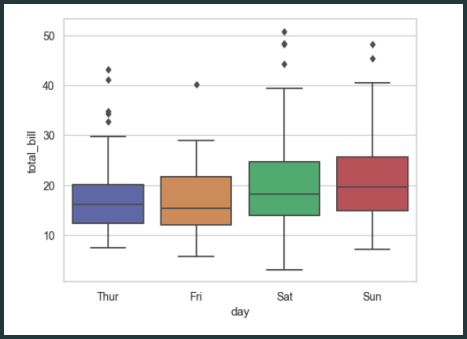
Box Plot

A plot introduced by William Tukey as a quick way to visualize the distribution of data.



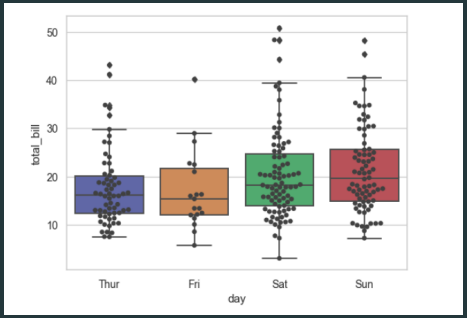
Code

```
1  import seaborn as sns
2  import pandas as pd sns.set_theme(style="whit
3  tips = sns.load_dataset("tips")
4  ax = sns.boxplot(x=tips["total_bill"])
```



Code

```
1  import seaborn as sns
2  import pandas as pd sns.set_theme(style="whit
3  tips = sns.load_dataset("tips")
4  ax = sns.boxplot(x="day", y="total_bill", dat
```



Code

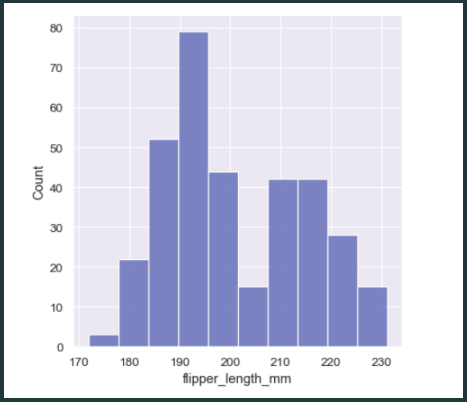
```
1  import seaborn as sns
2  import pandas as pd sns.set_theme(style="whit
3  tips = sns.load_dataset("tips")
4  ax = sns.boxplot(x="day", y="total_bill", dat
5  ax = sns.swarmplot(x="day", y="total_bill", d
```

Frequency Table

A table of the count of numeric data values that fall into a set of intervals (bins).

Histogram

A plot of the frequency table with the bins on the x-axis and the count (or proportion) on the y-axis.

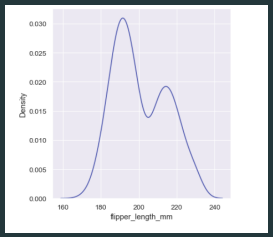


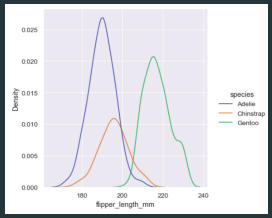
Code

```
1  import seaborn as sns
2  import pandas as pd
3  sns.set_theme(style="whitegrid")
4  penguins = sns.load_dataset("penguins")
5  sns.displot(penguins, x="flipper_length_mm",
```

Density Plot

A smoothed version of the histogram, often based on a kernal density estimate.





Code

```python
import seaborn as sns
import pandas as pd
sns.set_theme(style="whitegrid")
penguins = sns.load_dataset("penguins")
sns.displot(penguins,
x="flipper_length_mm", kind="kde")
```

```python
sns.displot(penguins, x="flipper_length_mm",
hue="species", kind="kde")
```

Modality

Unimodality

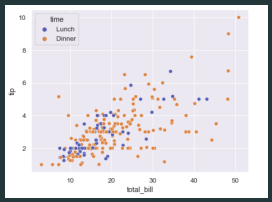
Unimodality means there is only a single highest value. (One hilltop)

Bimodality

Bimodality means there is two locally highest values. (Two hilltops)

Scatter Plot

A type of plot where the data are displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of the other variable determining the position on the vertical axis.



Code

```
    import seaborn as sns
    import pandas as pd
    tips = sns.load_dataset("tips") sns.scatt
x="total_bill", y="tip",
```

```
sns.displot(penguins, x="flipper_length_mm",
hue="species", kind="kde")
```

Thank you for your attention!

Appendix