University of Lancashire

Data Analysis using Visualisations

Karen O'Shea

Where opportunity creates success

Starter...

University of Lancashire

Why do we need to visualise data for example, how could this benefit us in the early stages of analysis?

What insights could you gain from using visuals (graphs/charts)?

Visualisation

Record information:

- Analyse data to support reasoning
  - Develop hypotheses
  - Find patterns
  - Discover errors
- Communicate to different audiences
  - Share, persuade and collaborate
- Psychology
  - How do people perceive and comprehend visual information?
  - Develop principles for creating effective visualisations

Learning Objectives

University of Lancashire

Prepare data for appropriate visualisation

Evaluate a dataset for quality control

Recap:
Data Cleaning…

Data Analysis?

- Only as good as your data
- Garage data in, garbage data analysis out

- Quality data decision-making!

Recap: So What is Data Cleaning?

- The process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset
- If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct
- There is no one absolute way to prescribe the exact steps in the data cleaning process
- Establish a method for good practice

Recap: Remove Unwanted/Duplicate Values

University of Lancashire

Remove unwanted observations from your dataset, including duplicate observations or irrelevant observations.

- Duplicate observations will happen most often during data collection.
- May occur when you combine data sets from multiple places, scrape data, or receive data from clients or multiple departments.

Irrelevant observations are when you notice observations that do not fit into the specific problem you are trying to analyse.

- For example, if you want to analyse data regarding millennial customers, but your dataset includes different generations, you might remove those irrelevant observations.

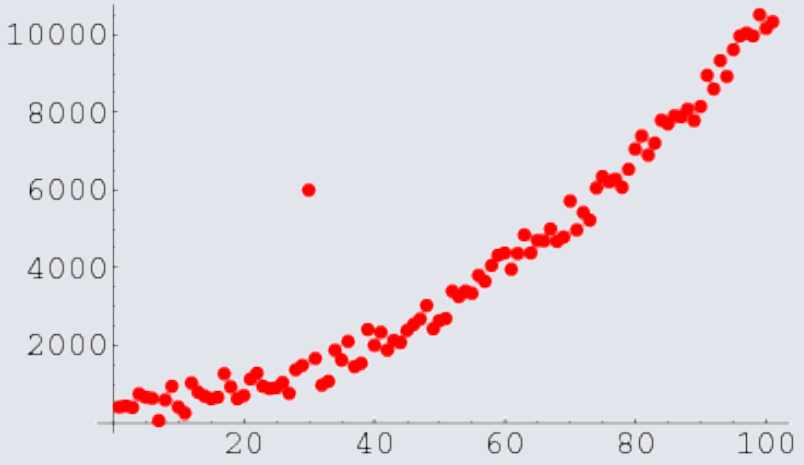
Recap: Fix Structural Errors

- Structural errors are when you measure or transfer data and notice strange naming conventions, typos, or incorrect capitalisation. These inconsistencies can cause mislabelled categories or classes:
  - For example, you may find “N/A” and “Not Applicable” both appear, but they should be analysed as the same category.

Recap: Filter Unwanted Outliers

- Often, there will be one-off observations where, at a glance, they do not appear to fit within the data you are analysing.
  - If you have a legitimate reason to remove an outlier, like improper data-entry, doing so will help the performance of the data you are working with.

- Sometimes it is the appearance of an outlier that will prove a theory you are working on.
  - Remember: just because an outlier exists, doesn't mean it is incorrect. If an outlier proves to be irrelevant for analysis or is a mistake, consider removing it.



Recap: Handle Missing Data

University of Lancashire

You can’t ignore missing data because many algorithms will not accept missing values.

- **As a first option**: you can drop observations that have missing values, but doing this will drop or lose information, so be mindful of this before you remove it.
- **As a second option**: you can input missing values based on other observations; again, there is an opportunity to lose integrity of the data because you may be operating from assumptions and not actual observations.
- **As a third option**: you might alter the way the data is used to effectively navigate null values.

Recap: Quality Control

University of Lancashire

Validate:

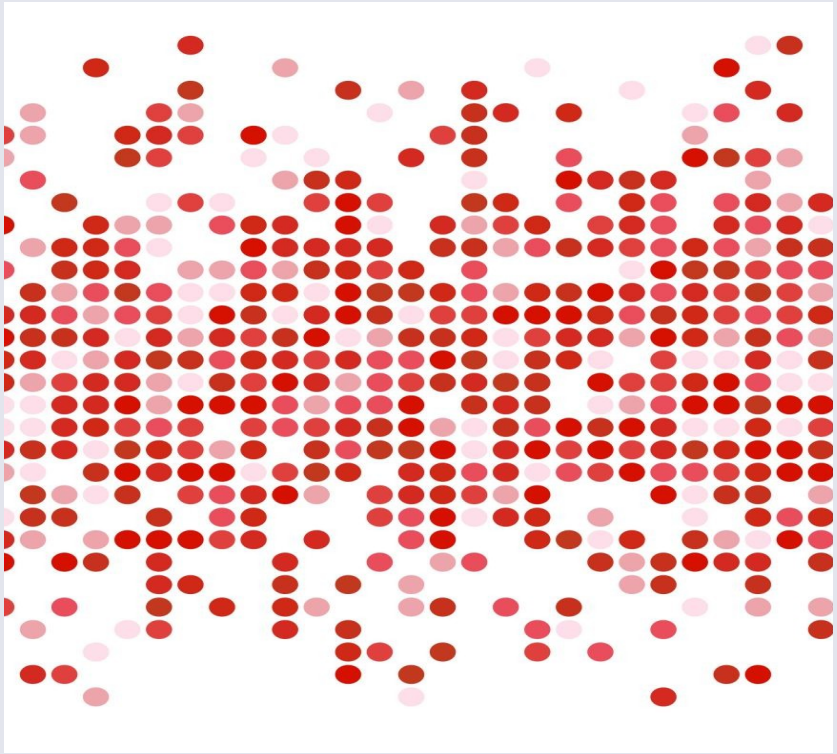
- Does the data make sense?
- Does the data follow the appropriate rules for its field?
- Does it prove or disprove your working theory, or bring any insight to light?
- Can you find trends in the data to help you form your next theory?
- If not, is that because of a data quality issue?

False conclusions because of incorrect or “dirty” data can inform poor business strategy and decision-making. Does your data stand up to scrutiny.

* What does data quality mean to you?

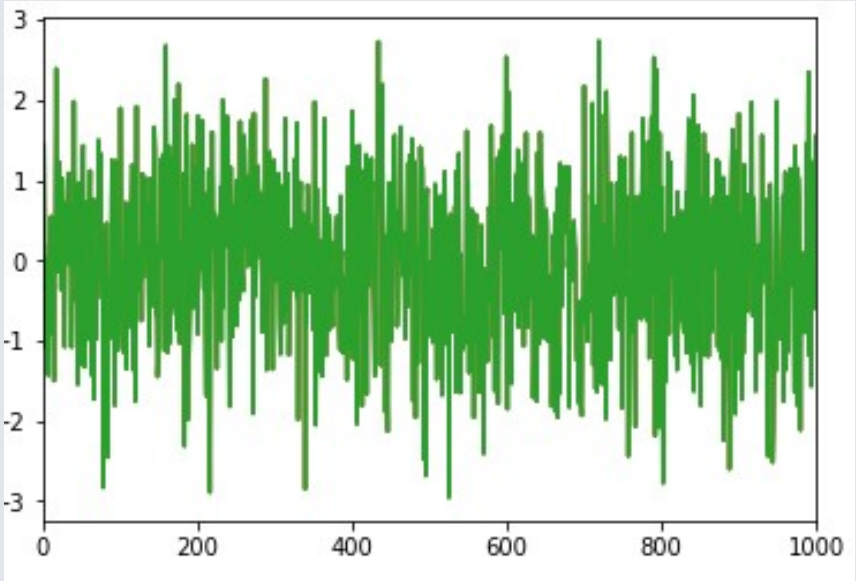
Noisy Data

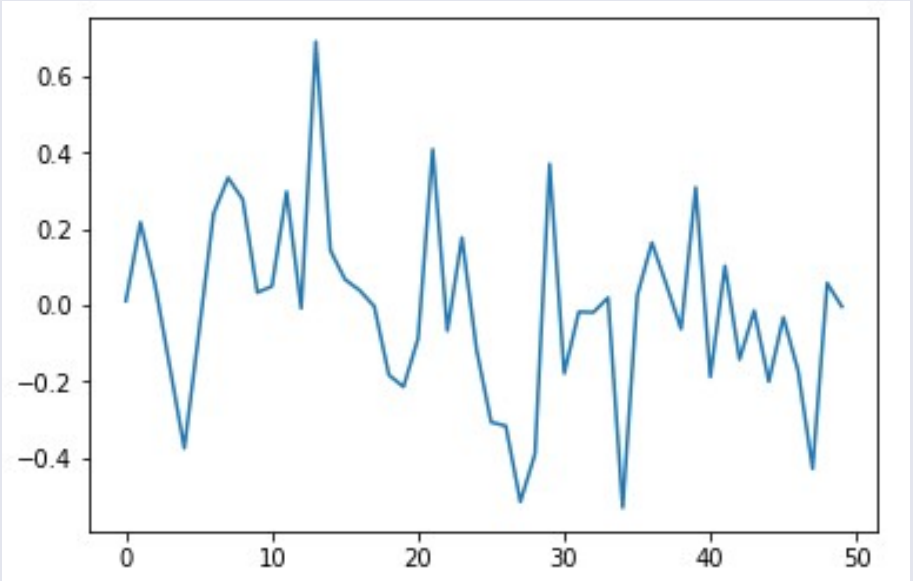
University of Lancashire



Why choose appropriate visuals?

University of Lancashire





Examine Data: Visualisations

- Prepare for Data Analysis
- Quick insights, outliers and patterns
  - Bar charts and Histograms – see patterns immediately
  - Line chart – identify if a trend is rising
  - Pie chart – view the magnitude of a factor compared to others

Further examples:

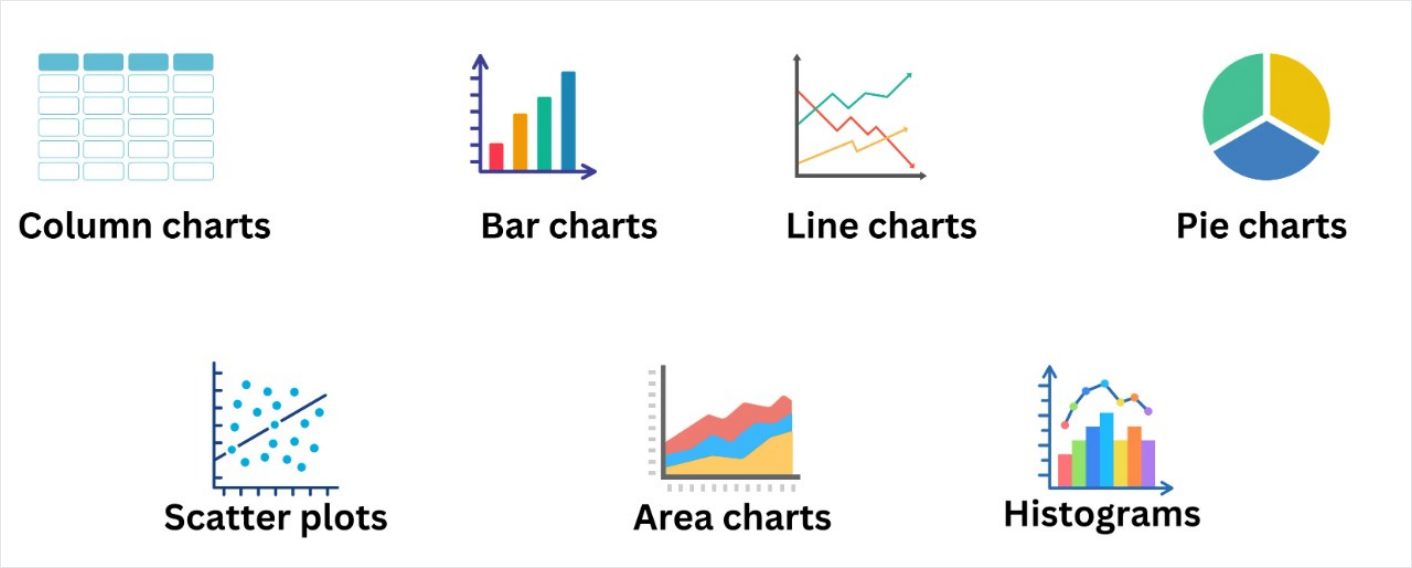
- Colour maps, story telling, network designing, explorative design, data models

https://trends.google.com/trends/explore?q=bitcoin

- Imagine that presented in a list of numbers

Statistical examples

- Start with basic plots – Line, Bar chart, Histogram and Scatter plot
  - Examine distributions, vary parameters/intervals.



Nature of the data

University of Lancashire

Data types

Quantitative

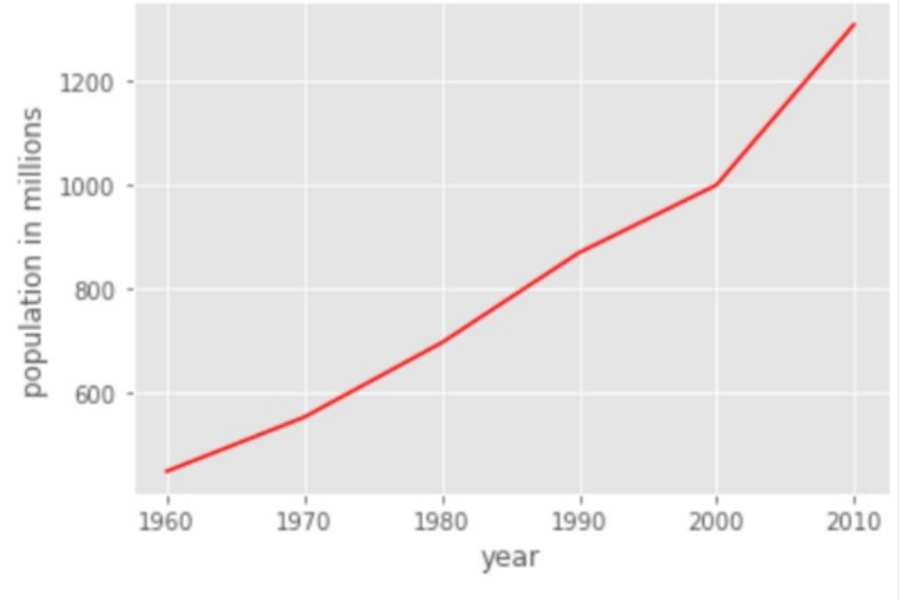
- Discreate data – numerical, finite number. For example, the number of employees in an office
- Continuous data – can take any value. For example, height, weight or time

Qualitative

- Categorical – quality or characteristic
  - Nominal – without rank or order. For example, eye colour, type of car, marital status
  - Ordinal – natural order or rank. For example, satisfaction ratings, (first, second, third; small, medium, large)

Line Chart with ‘Style’ Package

```
style.use(‘ggplot’)
year = [1960, 1970, 1980, 1990, 2000, 2010]
population = [449.48, 553.57, 696.783, 870.133, 1000.4, 1309.1]
plt.plot(year, population, color=‘red’)
plt.xlabel(‘year’)
plt.ylabel(‘population in millions’)
plt.title(‘population up to 2010’)
plt.show()
```



*** Note: style package. Import as follows:**

```
import matplotlib.style as style
```

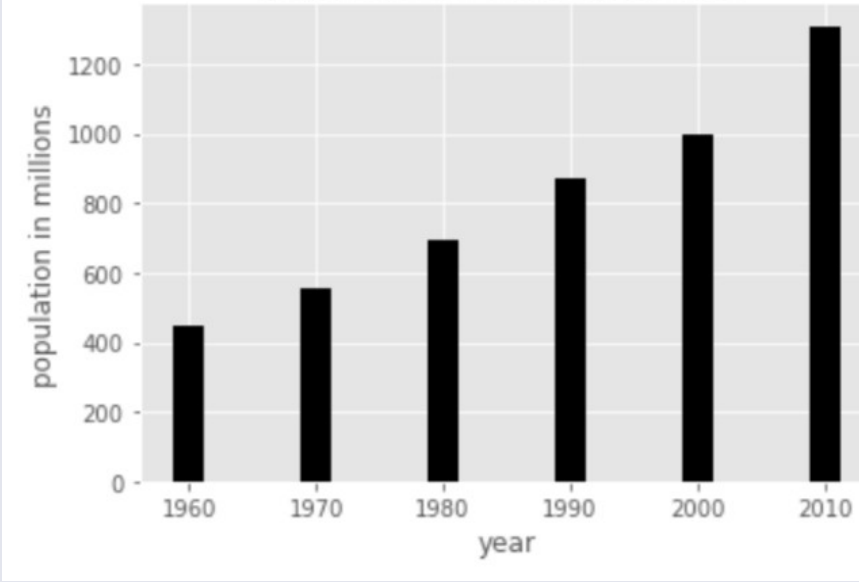
Bar Charts

University of Lancashire

```
bar_width=2.5

year = [1960, 1970, 1980, 1990, 2000, 2010]
population = [449.48, 553.57, 696.783, 870.133, 1000.4, 1309.1]

plt.bar(year, population, bar_width, color='black')
plt.xlabel('year')
plt.ylabel('population in millions')
plt.title('population up to 2010')
plt.show()
```



Multiple Lines on One Line Chart

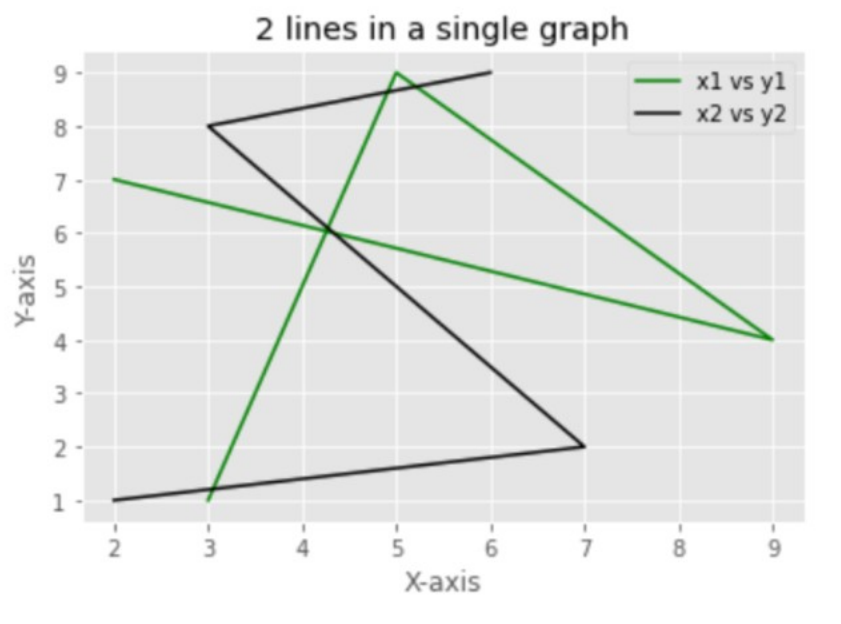
University of Lancashire

x1=[3, 5, 9, 2], y1=[1, 9, 4, 7]

x2=[6, 3, 7, 2], y2=[9, 8, 2, 1]

```
plt.plot(x1, y1, color='Green', label='x1 vs y1')
plt.legend()
plt.plot(x2, y2, color='Black', label='x2 vs y2')
plt.legend()

plt.xlabel('X-axis')
plt.ylabel('Y-axis')
plt.title('2 lines in a single graph')
plt.show()
```



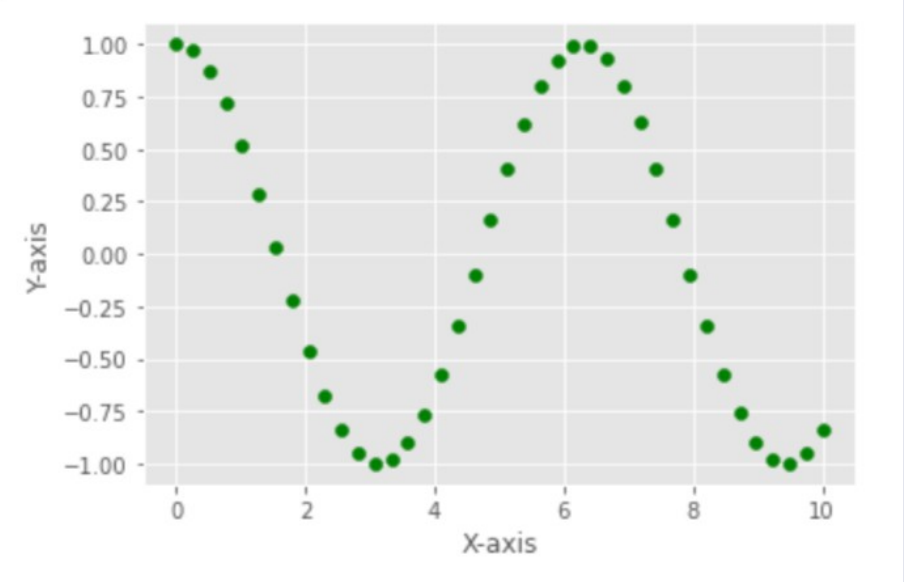
Scatter Plot

University of Lancashire

```
x=np.linspace(0, 10, 40)
y=np.cos(x)

plt.scatter(x, y, marker='o',
color='green’)

plt.xlabel('X-axis')
plt.ylabel('Y-axis')
plt.show()
```

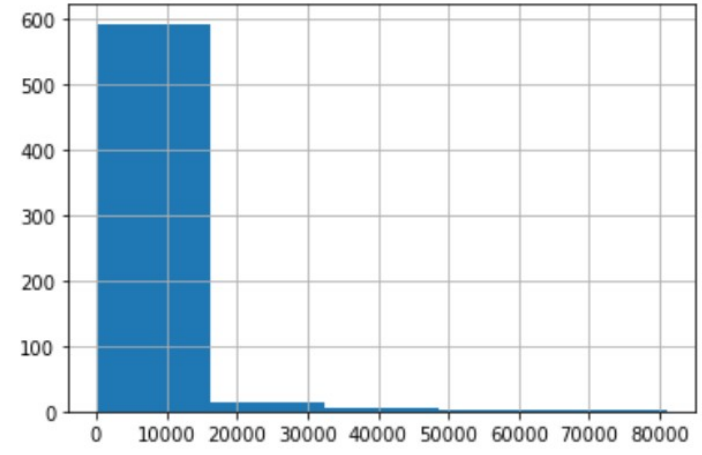


Comparison – Histogram ‘bins’ (intervals)
(Loan Prediction Dataset)

University of Lancashire

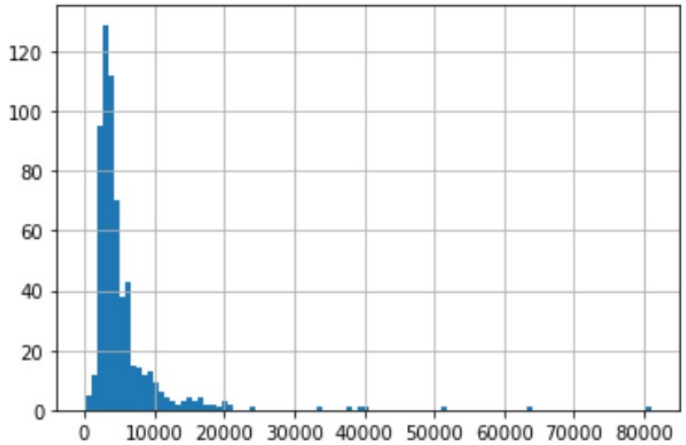
```
df['ApplicantIncome'].hist(bins=5)
```

```
<AxesSubplot:>
```




```
df['ApplicantIncome'].hist(bins=100)
```

```
<AxesSubplot:>
```



Hidden Trends – Smooth Data

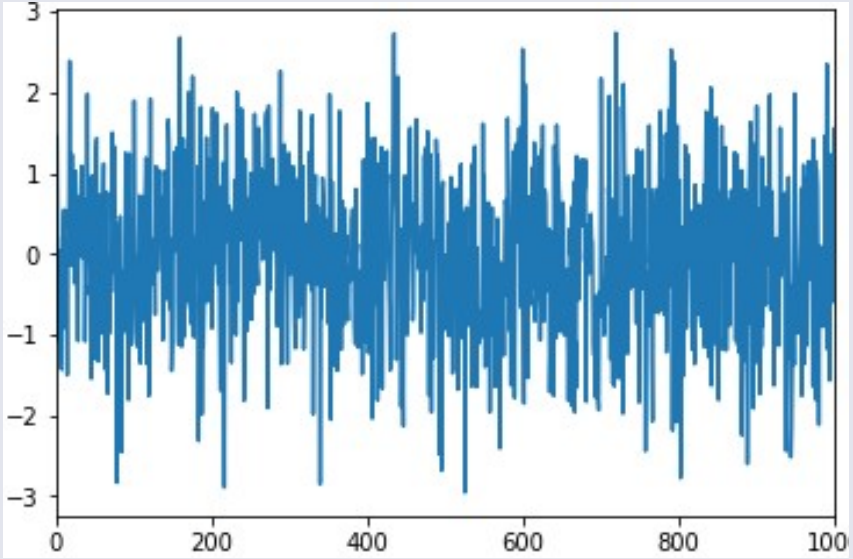
University of Lancashire

Smoothing by bin – each value in a bin is replaced by the mean value of the bin.

Smoothing by bin median – each bin value is replaced by its bin median value.

Smoothing by bin boundary – the minimum and maximum values in a given bin are identified as the bin boundaries.

* More on Statistical Analysis (Tools in Python) in later weeks



Basic rules for visualisation

- Follow formatting rules
  - Title, axis
- Context
  - Relate this to your questioning
  - Shopping patterns; climate change; fraud detection – changes in patterns
- Specific purpose and value
  - Has meaning
- Simplify complex data

Examine Data: Visualisations

- Using Matplotlib library:
  - **Installation — Matplotlib 3.4.3 documentation**

- *Matplotlib is a Python 2D plotting library that produces publication quality figures in various hardcopy formats and interactive environments across platforms….*

Learning Objectives

University of Lancashire

Prepare data for appropriate visualisation

Evaluate a dataset for quality control