GATE by moyyn

Data Analyst Module 2

Dr. Kavitha Chetana Didugu

BPW

Hochschule für Wirtschaft und Recht Berlin
Berlin School of Economics and Law
Startup Incubator Berlin
Das Gründungszentrum der HWR Berlin

STARTUP SCHOOL by Y Combinator

EUROPÄISCHE UNION
Europäischer Sozialfonds

be Berlin
Der Regierende Bürgermeister von Berlin
Senatskanzlei
Wissenschaft und Forschung

Today’s Agenda

Back to Basics

- Core Skill: Stats, Visualisation and Coding
- Descriptive Statistics and Visualisations
- Solving an example through code
- Introduction to Task 1

Descriptive and Inferential Statistics

Producing Insights and Recommendations

Analytical Tasks

Classification, Prediction, Association, Pattern Recognition

Classification

- Classification techniques helps in segmenting the customers into appropriate groups based on key characteristics.

- For example, using appropriate statistical model, an organization could easily segment the customers into Long term customers, medium term customers, and Brand switchers.

Pattern Recognition

- “A picture is worth thousand words” and it reveals hidden pattern in the data that could be leveraged be retail professionals. Pattern recognition techniques include *Histogram, Box Plot, Scatter plot and other visual analytics.*

- For example, histogram drawn for income of a particular class of customers may reveal a symmetrical bell curve pattern or may be left or right skewed.

Association

- *Association* analysis helps in determining which of the items go together. Association rules include a set of analytics that focuses on discovering relationships.

- In this context, market basket analysis refers to an association rule that generates the probability for an outcome.

Predictive Modeling

- Both customer segmentation as well as identifying and targeting most profitable customers can be facilitated by predictive models.

- Regression can be used for predicting the amount of expenditure on a particular product based on input variables income, age and gender.

Statistical Terminology

Types of statistics

Descriptive statistics is concerned with Data summarization, Graphs/Charts, and tables

Inferential statistics is a method used to talk about a Population parameter from a sample

- A population is the universe of possible data for a specified object.
- A parameter is a numerical value associated with a population.
- A sample is a selection of observations from a population
- A Statistic is a numerical value associated with an observed sample.

Data Types

Types of Data

- Qualitative data are non numeric in nature and cannot be measured.
- Quantitative data are numerical in nature and can be measured and can be classified into two: discrete and continuous.
  - Discrete type can take only certain values, and there are discontinuities between values
  - Continuous type can take any value within a specific interval.

Types of datasets

- Record
- Graph and network
- Ordered
- Spatial, image and Multimedia

Data attributes

Data objects

- Data sets are made up of data objects.
- A data object represents an entity.
- Data objects are described by attributes.

Examples:

- sale database : customers, sales
- medical database: patients, treatments

Attributes

Data field, representing a characteristic or feature of a data object

Example:

- customer_ID, name, address

Types:

- Nominal, Binary, Ordinal, Numeric

Descriptive Statistics

Important Concepts

- Frequency Distribution - histograms
- Cumulative frequency distribution
- Measures of central tendency
  - Mean, Median, Mode
- Measures of dispersion
  - Range, IQR, standard deviation, coefficient of variation
- Normal distribution
- Five number summary, boxplots, QQ plots
- Correlation analysis

Measures of Central Tendencies

Whenever you measure things of the same kind, a fairly large number of such measurements will tend to cluster around the middle value. Such a value is called a measure of “Central tendency”

Mean

The statistical mean refers to the mean or average that is used to derive the central tendency of the data.

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + \lambda + X_n}{n}$$

Measures of Central Tendencies

Median

The middle value that separates the higher half from the lower half of the data set. The median and the mode are the only measures of central tendency that can be used for original data, in which values are ranked relative to each other but are not measured absolutely.

Mode

The most frequent value in the data set. This is the only central tendency measure that can be used with nominal data, which have purely qualitative category assignments.

Measures of Dispersion, Range and IQR

Measure of dispersion indicate how large the spread of distribution in around the central tendency.

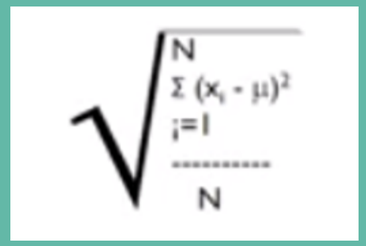
Range is the difference between maximum and minimum value in dataset.
Range = X(maximum) - X (minimum)

Inter-Quartile Range: Difference between third and first Quartile (Q3 - Q1)

Standard Deviation and Variance

•Interpreting variance (a squared term) is not intuitive. Instead we under root it to get Standard deviation which has the same units as variable.

•Standard deviation, is a measure of average spread i.e., on an average what is the difference between any data point and the central value of the variable.



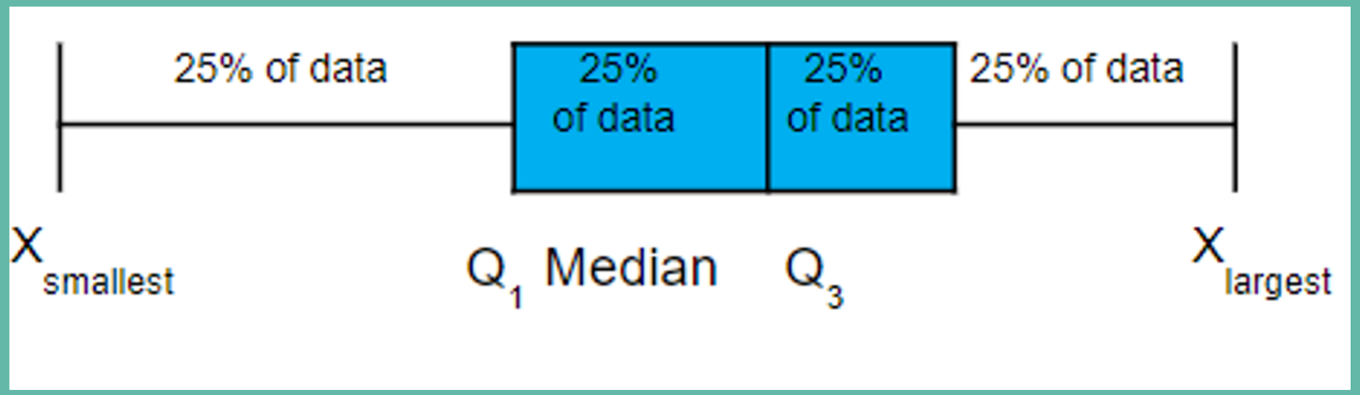
- Coefficient variation is defined as ratio of standard deviation to mean.

$$CV = \frac{S}{\overline{X}} \text{ for the sample data and } = \frac{\sigma}{\mu} \text{ for the population}$$

Five-Point Summary

The five numbers that help describe the center, spread and shape of data are:

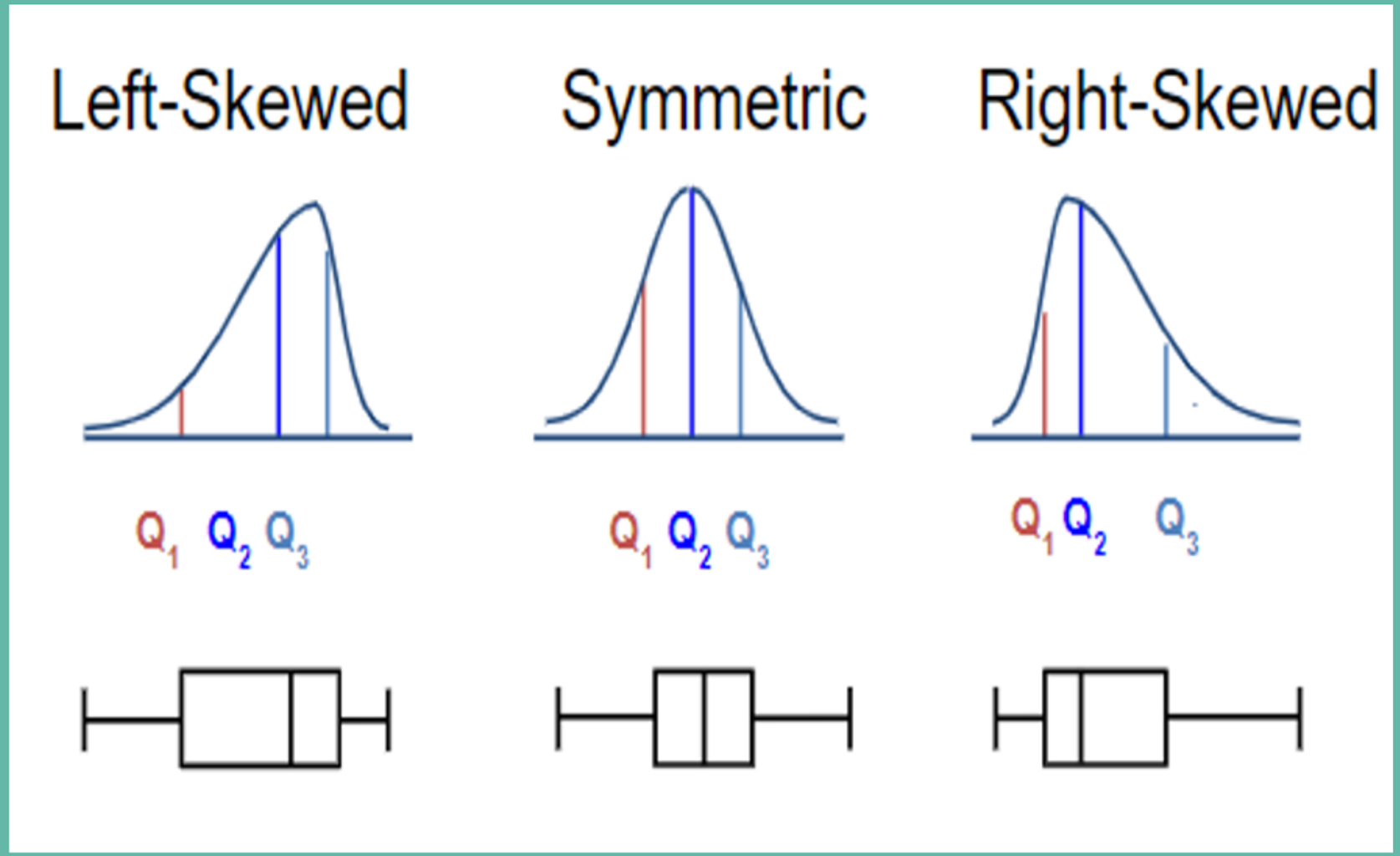
- **Xsmallest**
- **First Quartile (Q1)**
- **Median (Q2)**
- **Third Quartile (Q3)**
- **Xlargest**



The Boxplot: A Graphical display of the data based on the five-number summary

We saw the five point summary in Pandas describe() function!

Boxplot and Data Skewness

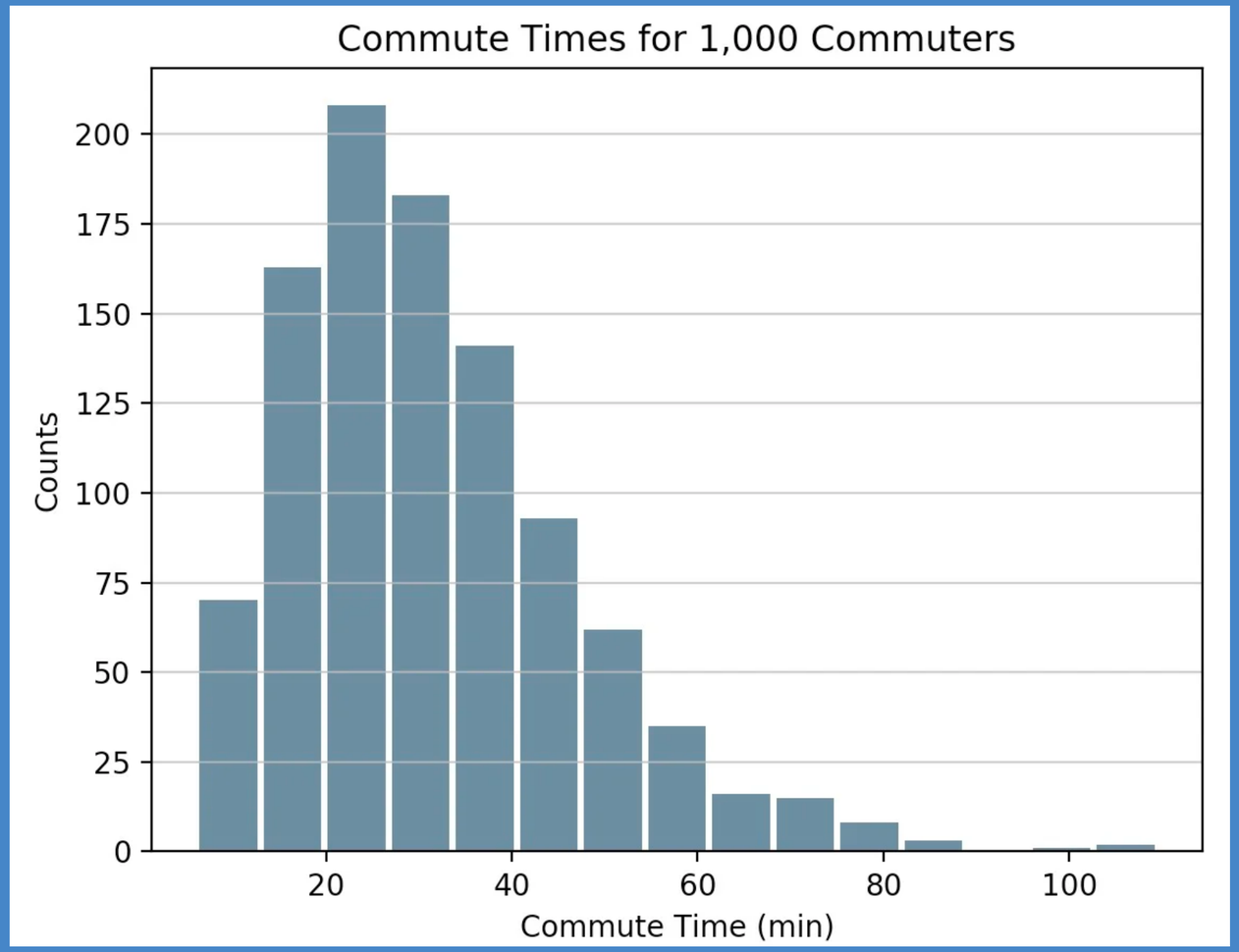


So many statistics, so many graphs

When to use which one?

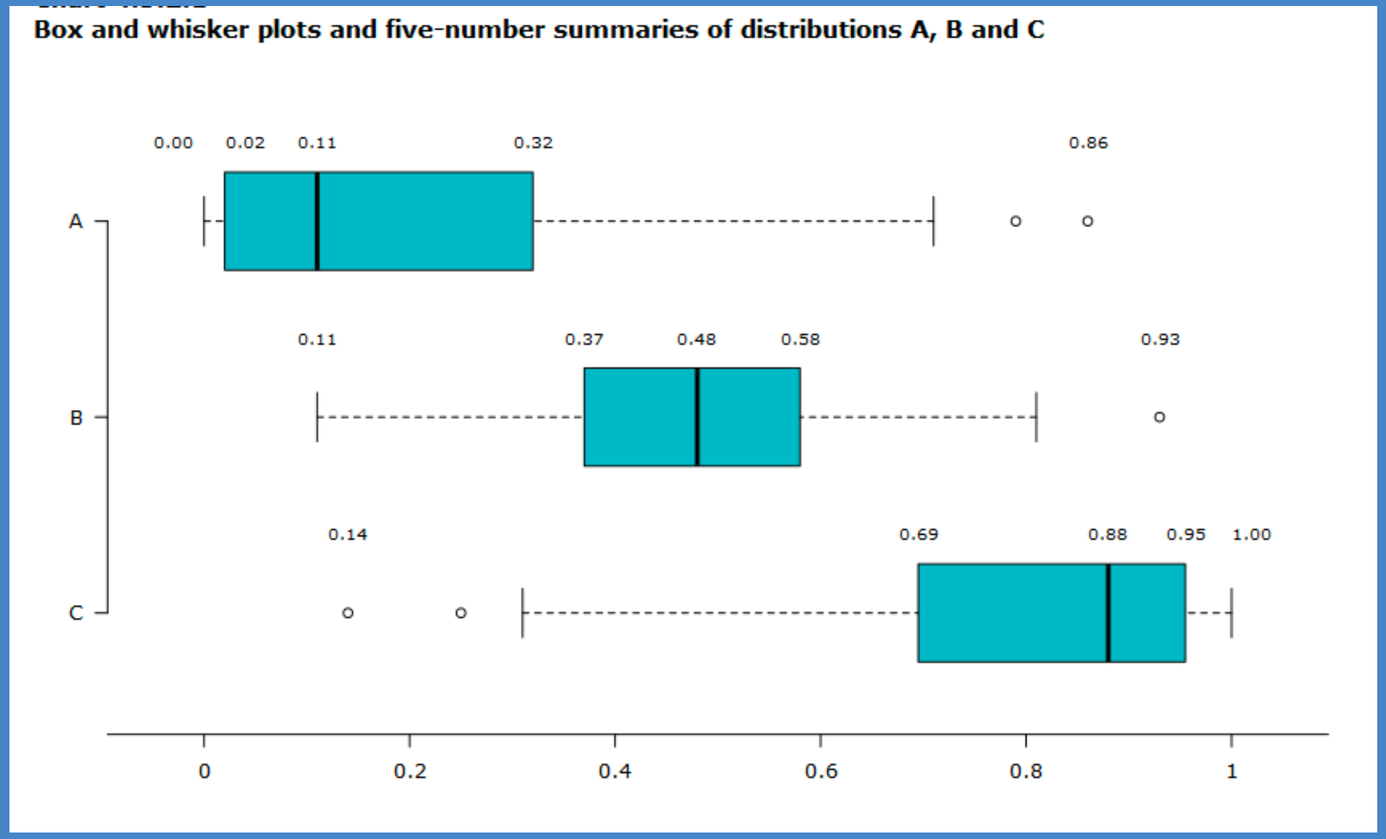
Some Visualisation Guidelines

- **Histogram:** x-axis are values/value ranges, y-axis shows frequencies



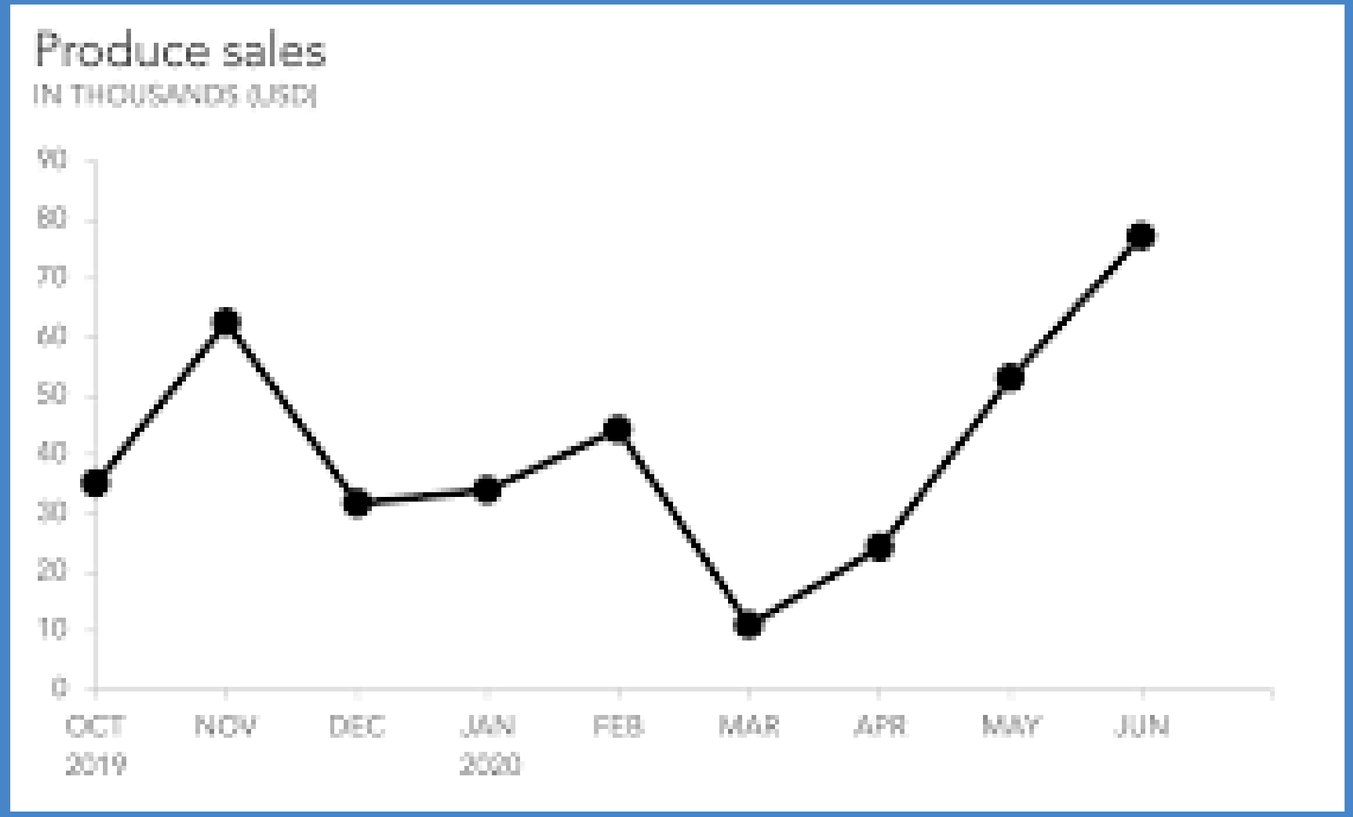
Some Visualisation Guidelines

- **Box plot:** when you want to identify skewness in a variable



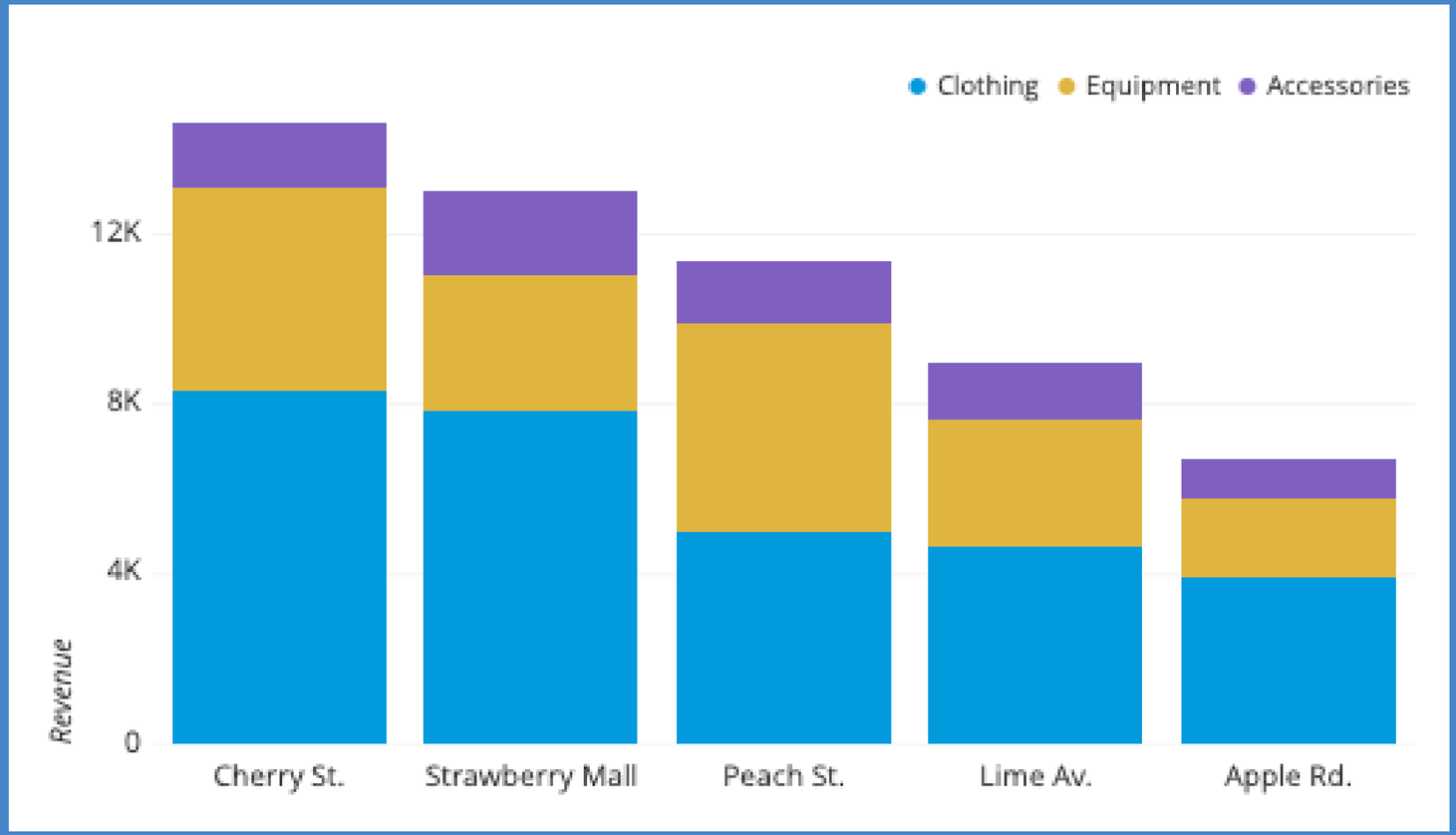
Some Visualisation Guidelines

- **Line plot:** When y axis is a numberical variable, and the x axis is a continuous variable like time



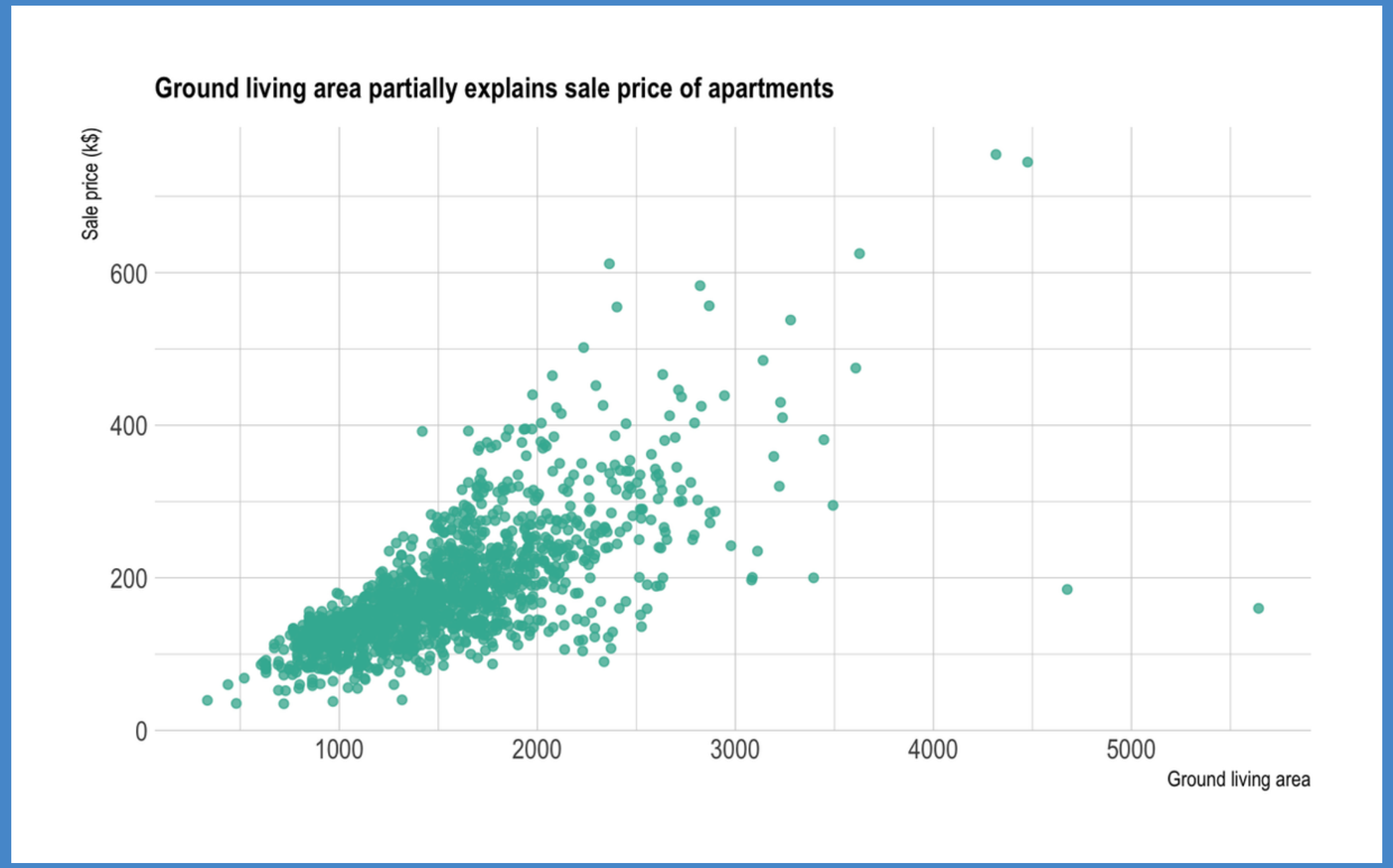
Some Visualisation Guidelines

- **Stacked plot:** to show multiple slices in a bar graph that make up the whole



Some Visualisation Guidelines

- **Scatter plot:** each pair of values is a pair of coordinates and plotted as points in the plane



Some Visualisation Guidelines

Bi-variate analysis: Correlation between two variables

$$r_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}}$$

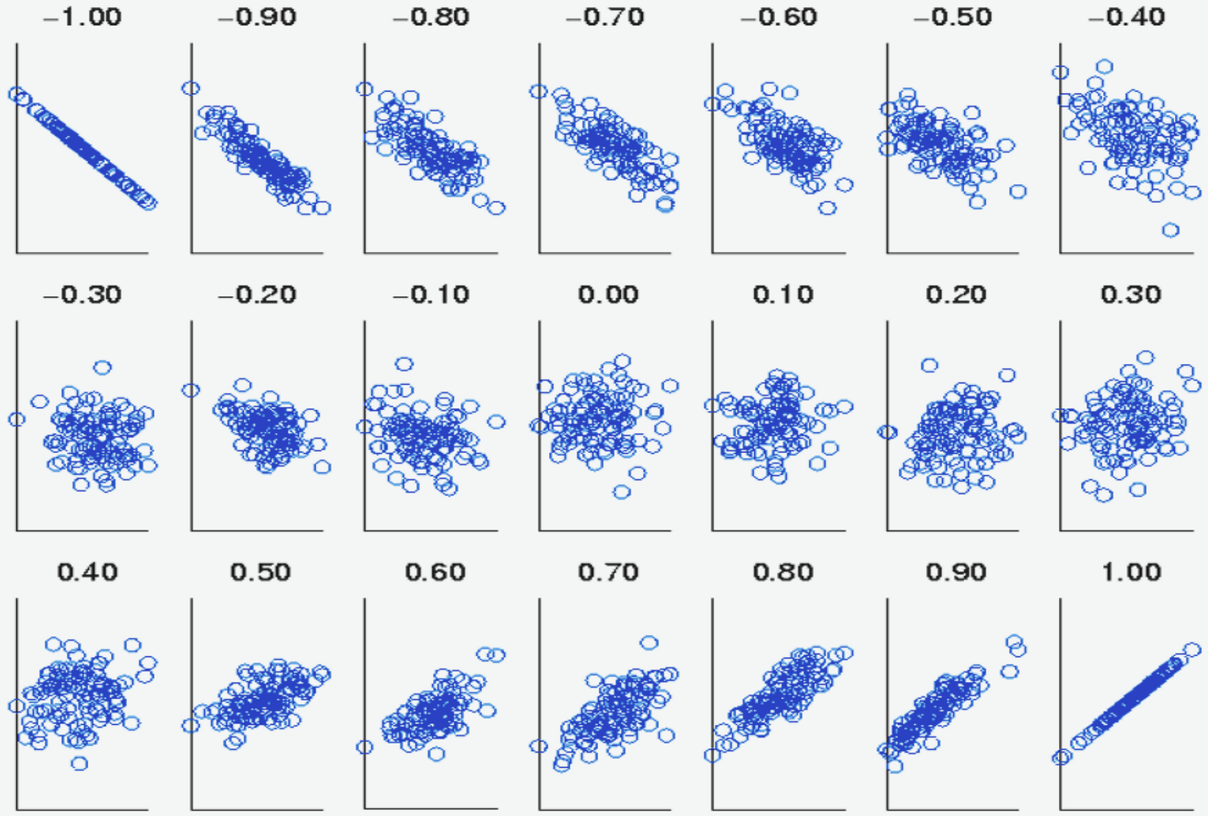
r_{xy} = correlation coefficient between x and y

x_i = the values of x within a sample

y_i = the values of y within a sample

$\bar{x}$ = the average of the values of x within a sample

$\bar{y}$ = the average of the values of y within a sample



Interaction

Task 1

Customer - Course Matching

You are given the customer engagement data for the website of an EdTech startup. Based on the data, provide recommendations on which type of customers tend to choose which course.

Link to the data:
https://www.kaggle.com/datasets/kavithachetanadidugu/crm-cleaned-data/settings

Feedback Survey

http://www.moyyn.com/gate-feedback

Further Reading and Resources

- http://rafalab.dfci.harvard.edu/dsbook/dataviz-distributions.html
- https://www.scribbr.com/statistics/inferential-statistics/
- https://www.csusm.edu/stemsc/handouts/stats_hypothesistesting.pdf
- https://blog.ml.cmu.edu/2020/08/31/7-causality/
- https://edu.gcfglobal.org/en/statistics-basic-concepts/sampling-methods/1/

GATE

by moyyn

German Academy for Technology and Entrpreneurship

BPW

Hochschule für Wirtschaft und Recht Berlin
Berlin School of Economics and Law

Startup Incubator Berlin
Das Gründungszentrum der HWR Berlin

STARTUP SCHOOL
by Y Combinator

EUROPÄISCHE UNION
Europäischer Sozialfonds

be Berlin

Der Regierende Bürgermeister von Berlin
Senatskanzlei
Wissenschaft und Forschung