Data Exploration

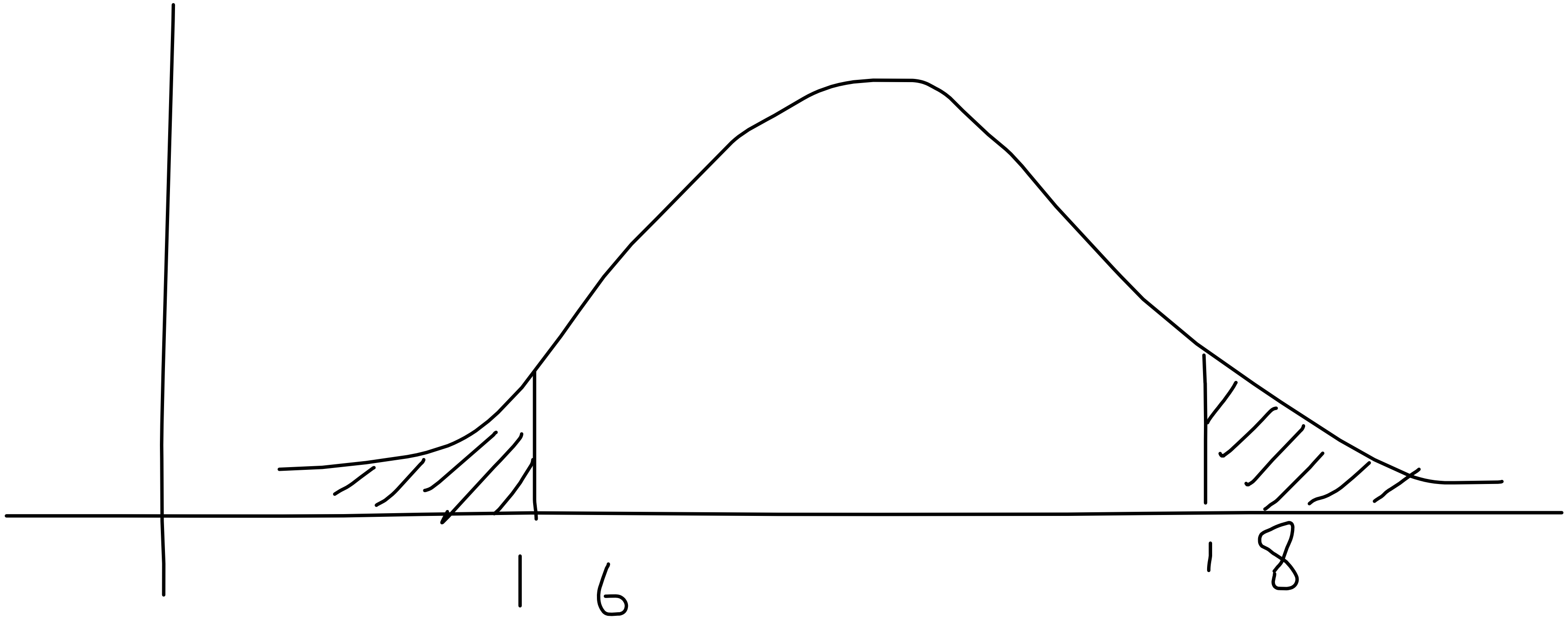
Goal of Data Exploration

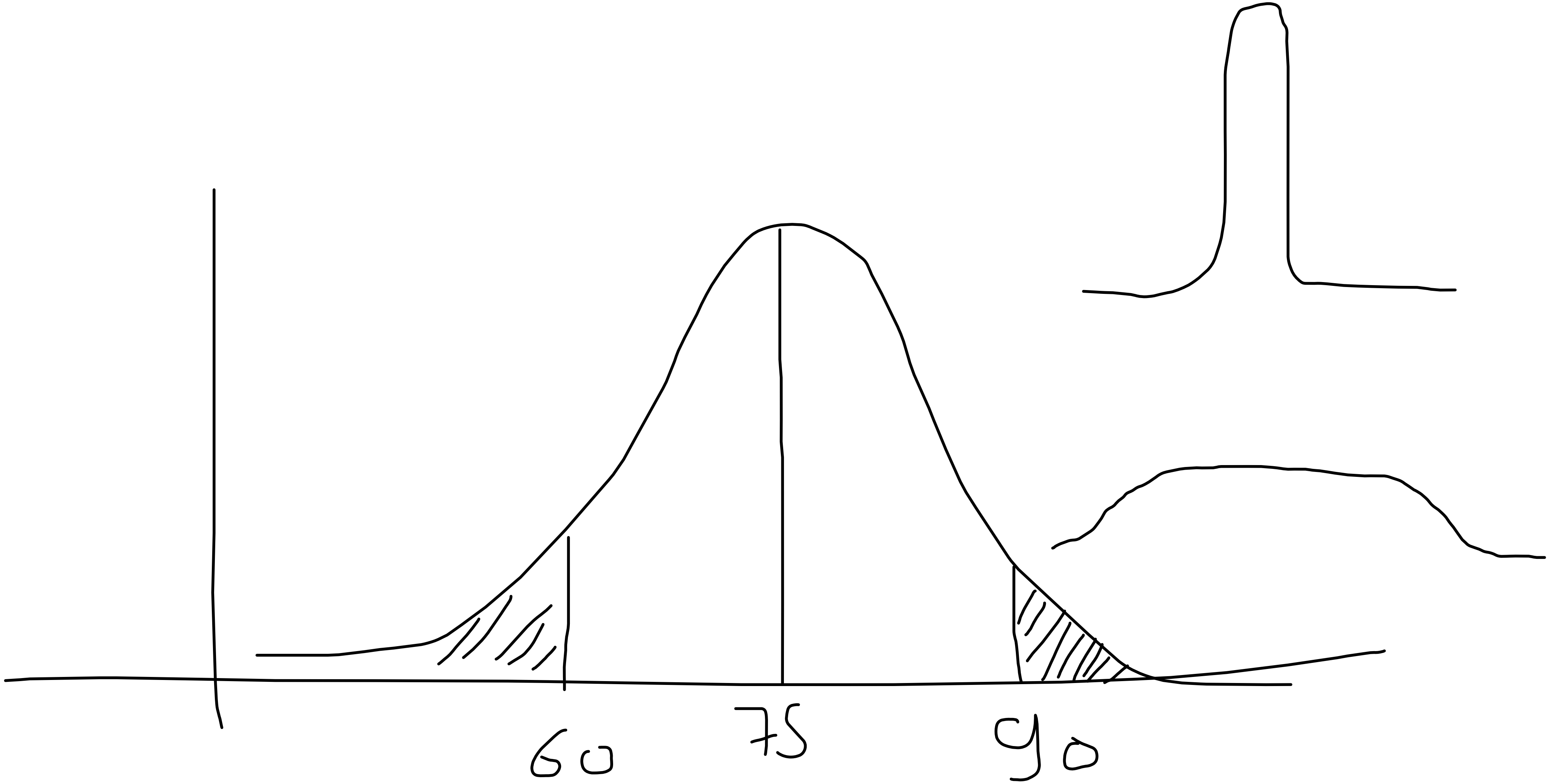
- Goal:
  - Understand the basic characteristics of the data

- Examples for characteristics:
  - Structure
  - Size
  - Completeness
  - Relationships

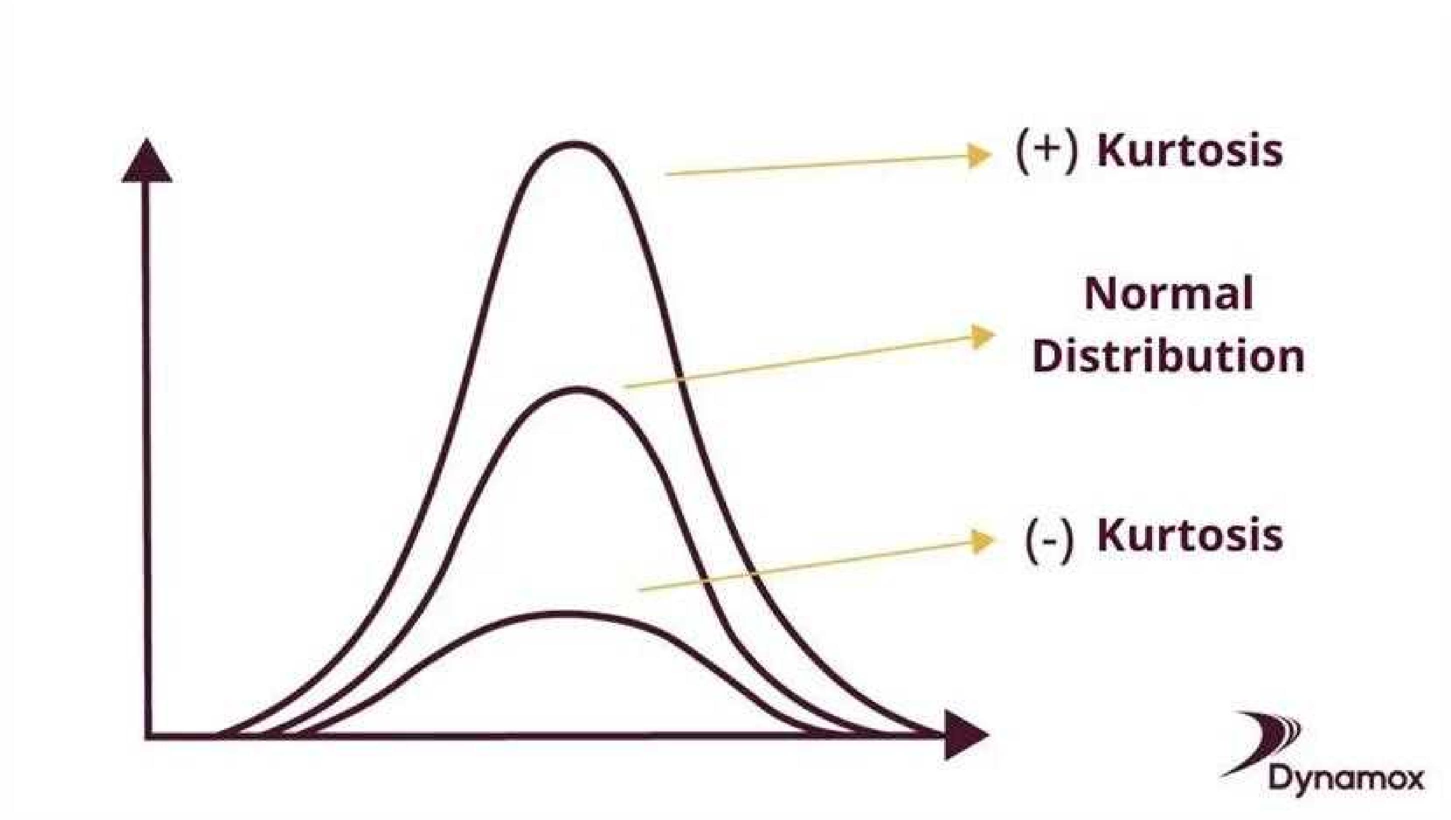
Descriptive Statistics

- Summarize data through single value
- Common statistics
  - Central tendency (mean/median/mode)
  - Variability (standard deviation, interquartile range)
  - Range of data (min/max)
- Other important statistics
  - Kurtosis and skewness for the shape of distributions

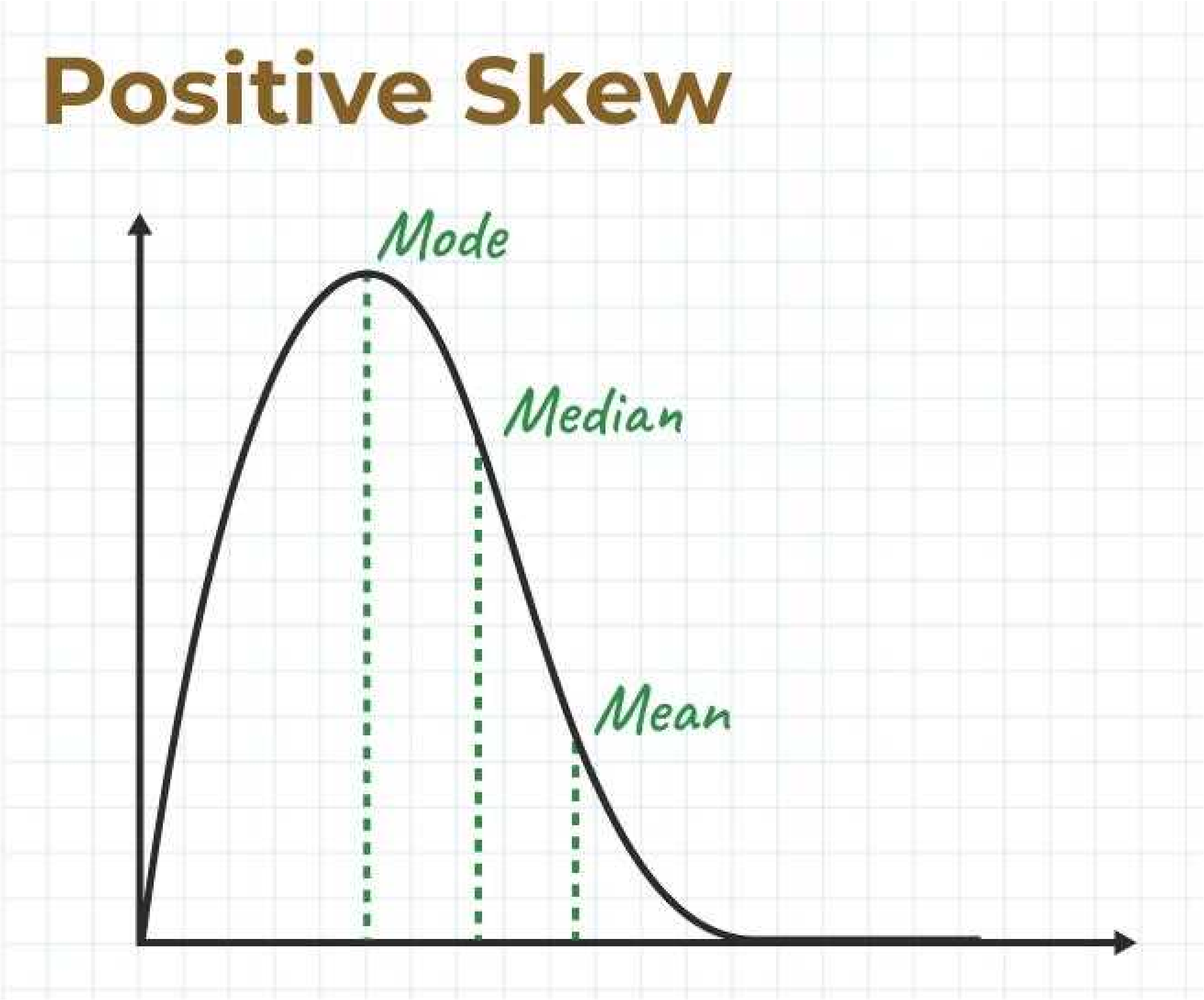

16
18


60
75
90

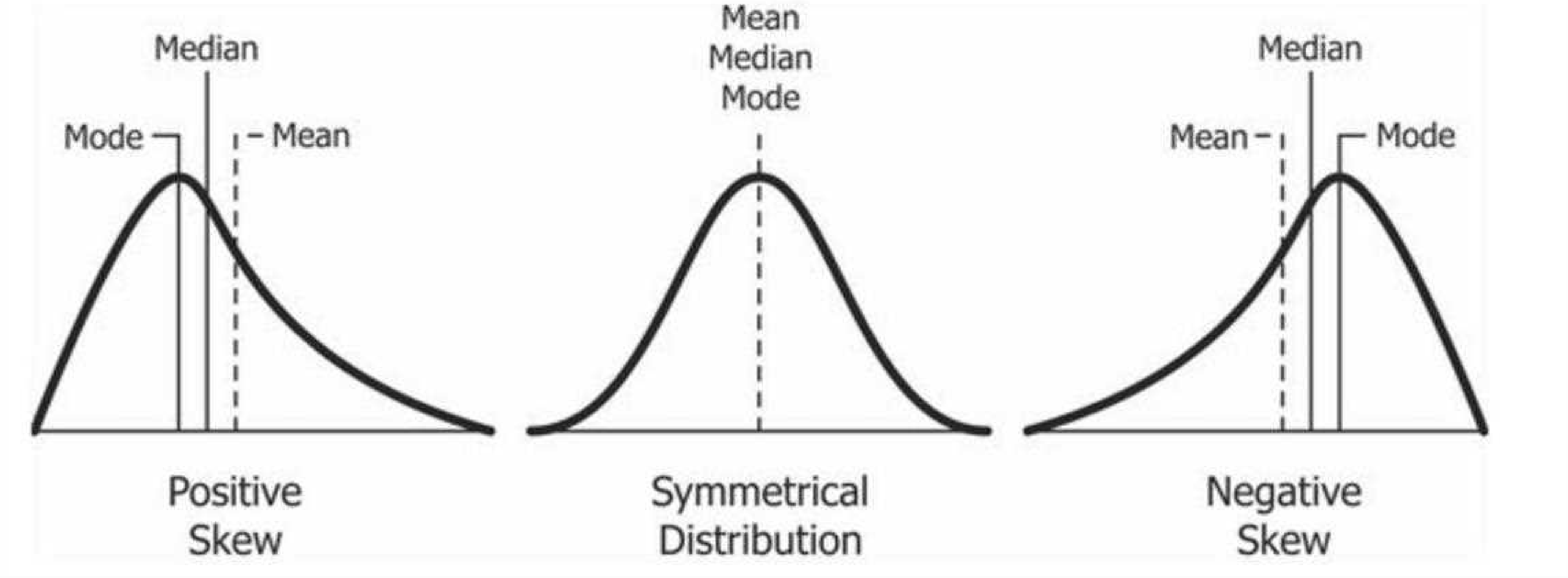
Kurtosis



skewness



skewness


Median
Mode
Mean
Positive Skew
Mean
Median
Mode
Symmetrical Distribution
Median
Mean
Mode
Negative Skew

Central Tendency

- Arithmetic mean
  - $mean(x) = \frac{1}{n}\sum_{i=1}^{n} x_i$ with $x = (x_1, \dots, x_n) \in \mathbb{R}^n$
- Median
  - The value that separates the higher half from the data of the lower half
- Mode
  - The value that appears most in the data

Variability

- Measure for the spread of the data

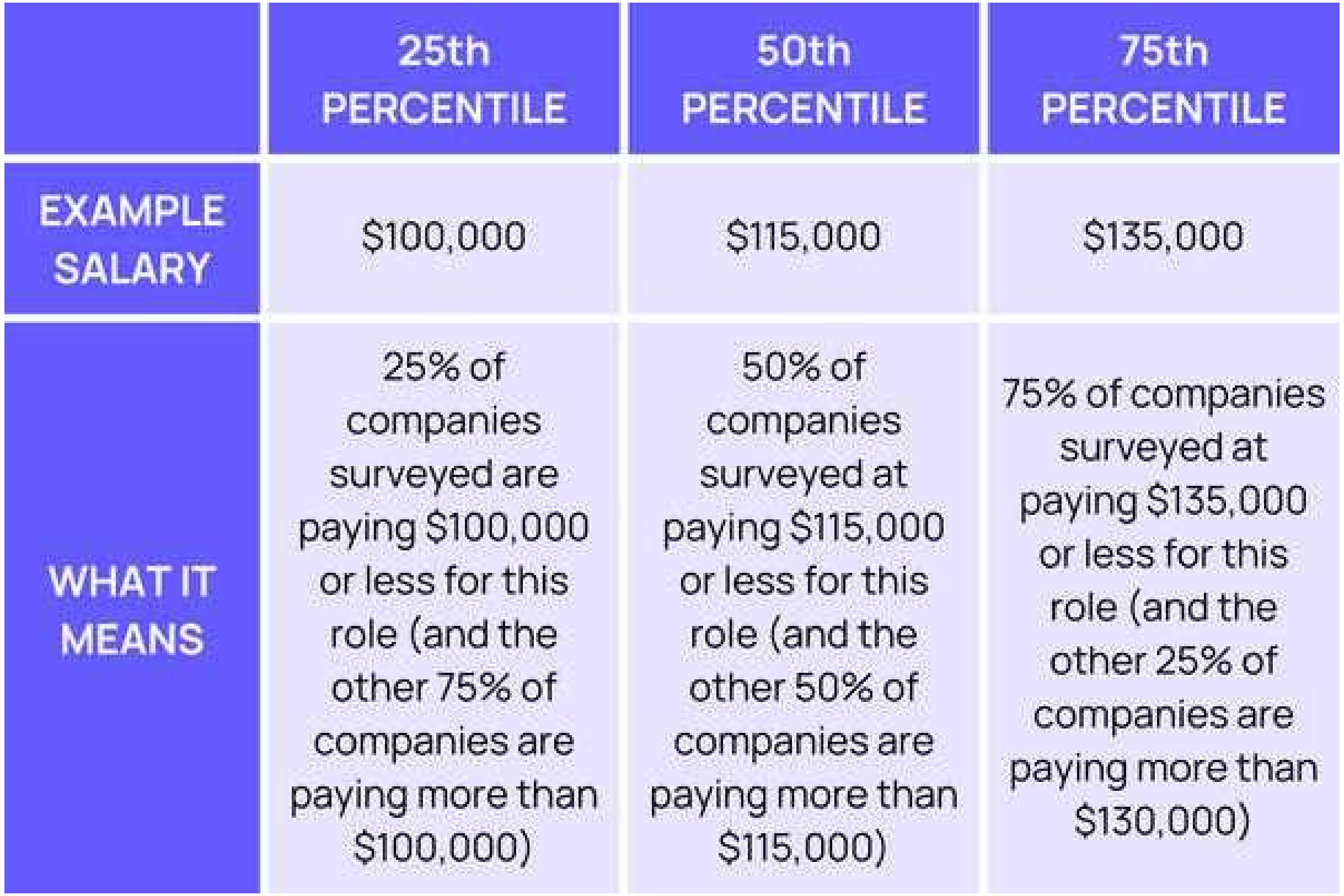
- Standard deviation
  - Measure for the difference of observation to the arithmetic mean
  - $sd(x) = \sqrt{\frac{\sum_{i=1}^{n} \left(x_i - mean(x)\right)^2}{n-1}}$

- Interquartile Range (IQR)
  - Percentile: value below which a given percentage falls
  - Difference between the 75% percentile and the 25% percentile

percentile

- K-th percentile is x
  - K% of the values are less than x
  - (100 – K) % of the values are larger than x

percentile



	25th PERCENTILE	50th PERCENTILE	75th PERCENTILE
EXAMPLE SALARY	$100,000	$115,000	$135,000
WHAT IT MEANS	25% of companies surveyed are paying $100,000 or less for this role (and the other 75% of companies are paying more than $100,000)	50% of companies surveyed at paying $115,000 or less for this role (and the other 50% of companies are paying more than $115,000)	75% of companies surveyed at paying $135,000 or less for this role (and the other 25% of companies are paying more than $130,000)

Range of data

- Range for which values are observed
- Minimum: Smallest observed value
- Maximum: Largest observed value

- May be strongly distorted by invalid data
  - Makes it also a good tool to discover invalid data

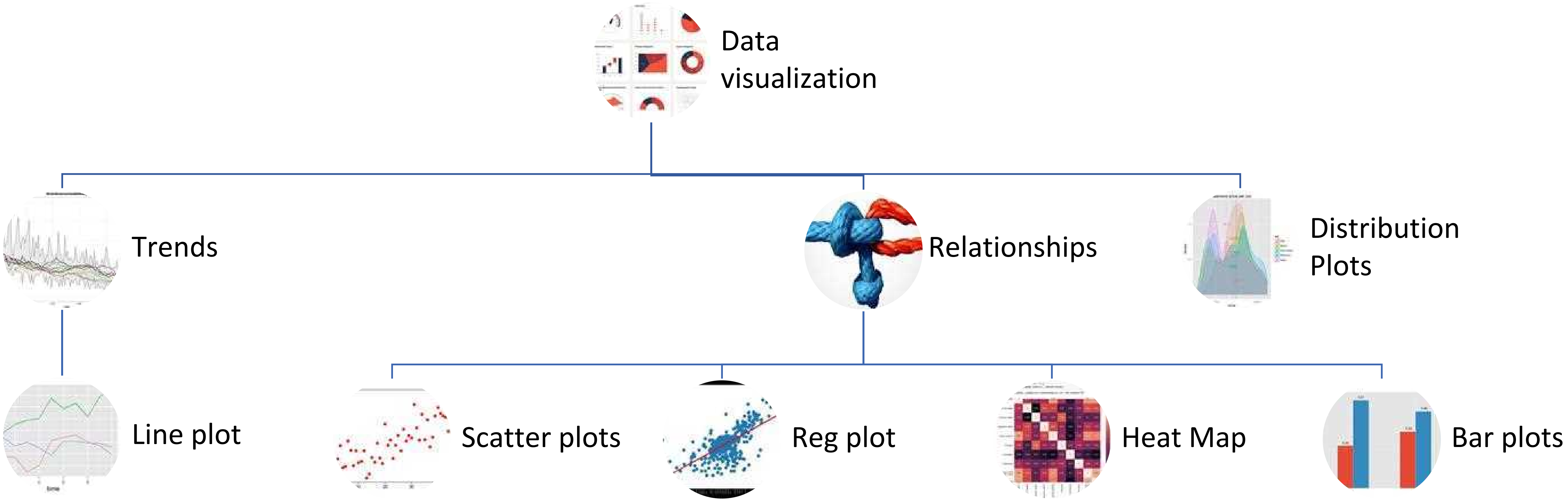
Example

- Random typing on the keypad
- x = (1, 2, 1, 1, 3, 4, 5, 2, 3, 4, 5, 1, 3, 2, 1, 6, 5, 4, 9, 4, 3, 6, 1, 5, 6, 8, 4, 6, 5, 1, 3, 2, 1, 6, 8, 7, 6, 1, 3, 1, 6, 8, 4, 7, 6, 4, 3, 5, 4, 9, 7, 4, 3, 1, 4, 6, 8, 7, 9, 1, 4, 6, 1, 3, 8, 6, 7, 4, 9, 6, 5, 1, 3, 6, 8, 7)
- central tendency:
    - mean: 4.46052631579
    - median: 4.0
    - mode (count): 1 (14)
- variability
    - sd: 2.41944311488
- range
    - min: 1
    - max: 9

Visualization

- A Picture Says More than 1000 Words
- Processed faster than textual information
- Simplify complex information
- Show patterns
- Better long memrization (information retention)

Types of visualizations



Data visualization

Trends

Relationships

Distribution Plots

Line plot

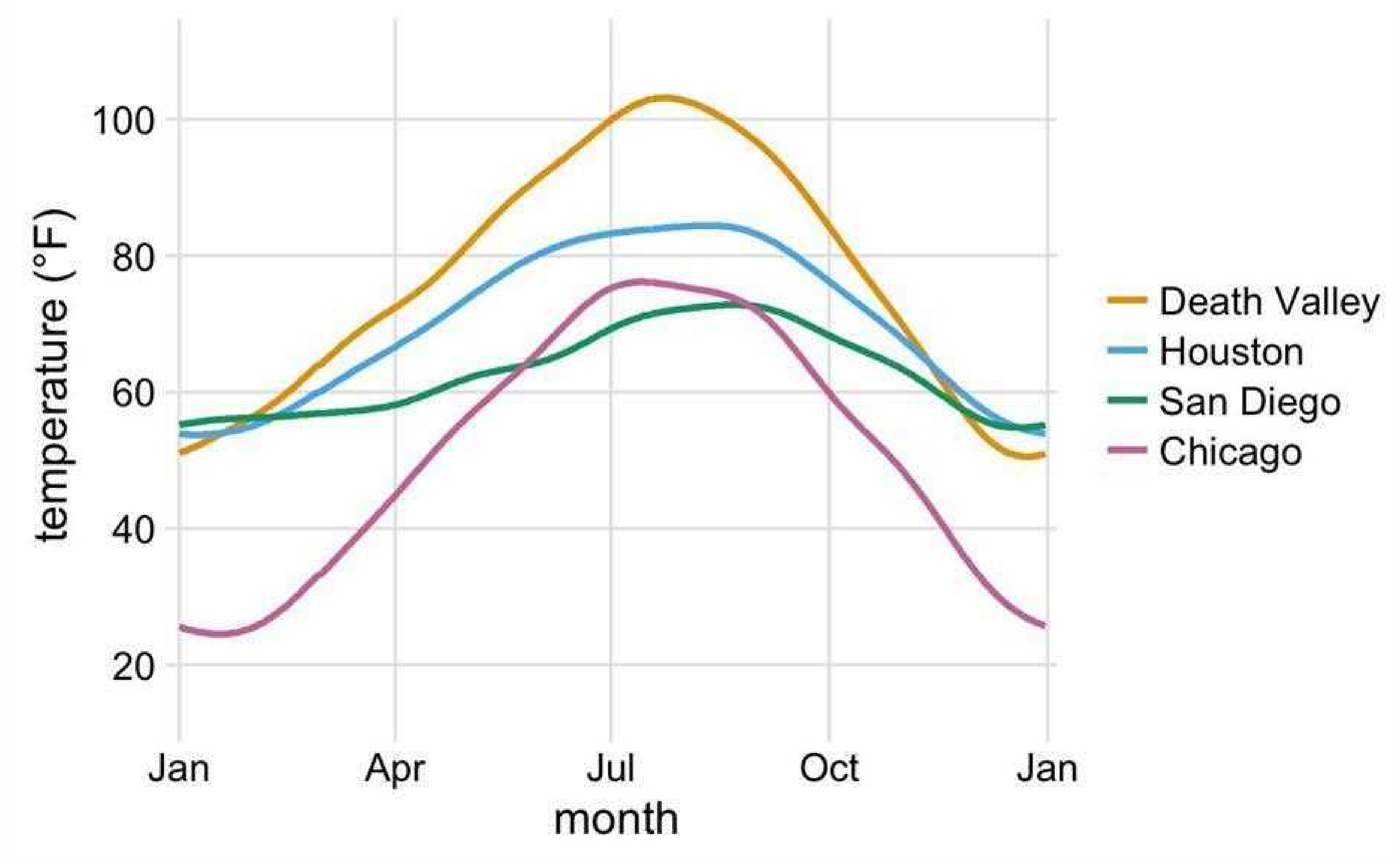
Scatter plots

Reg plot

Heat Map

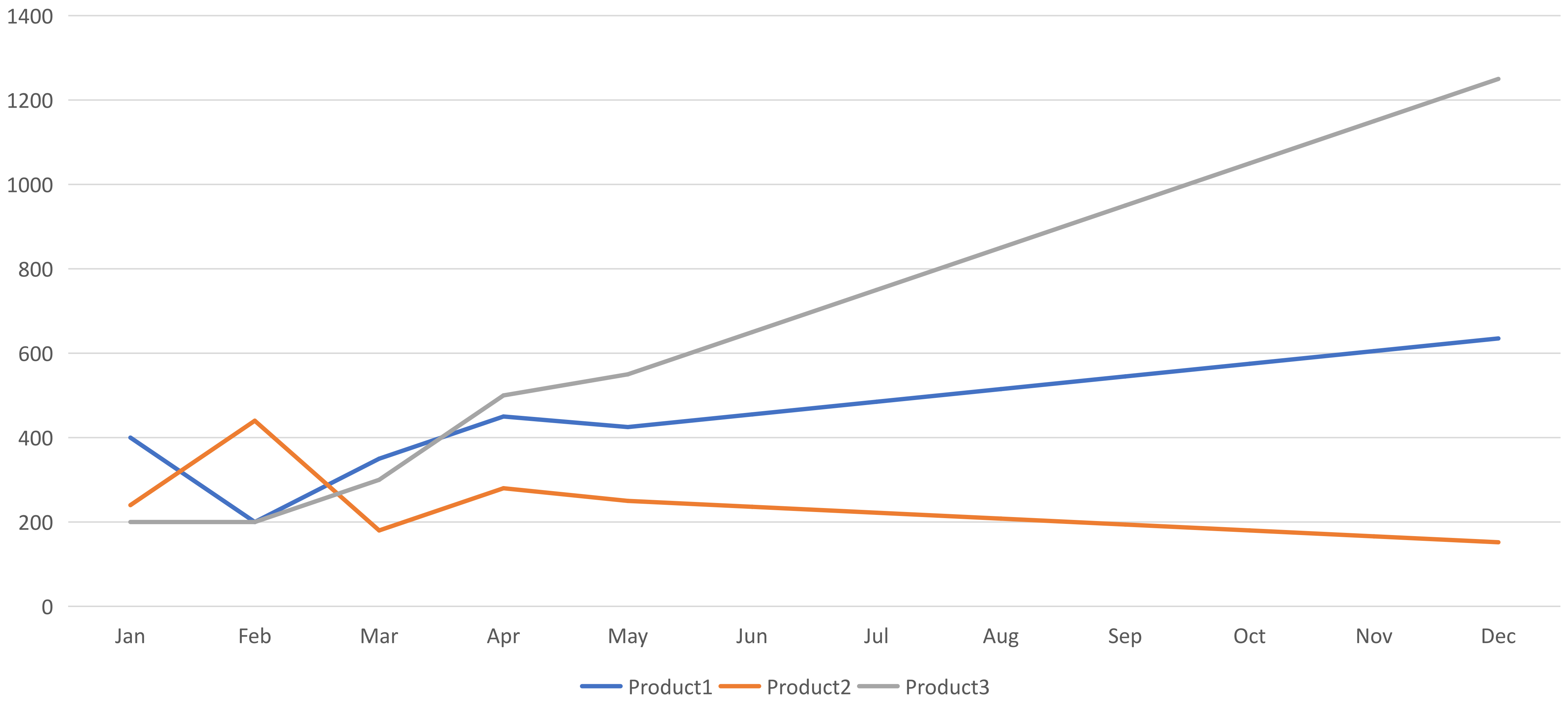
Bar plots

Line plot

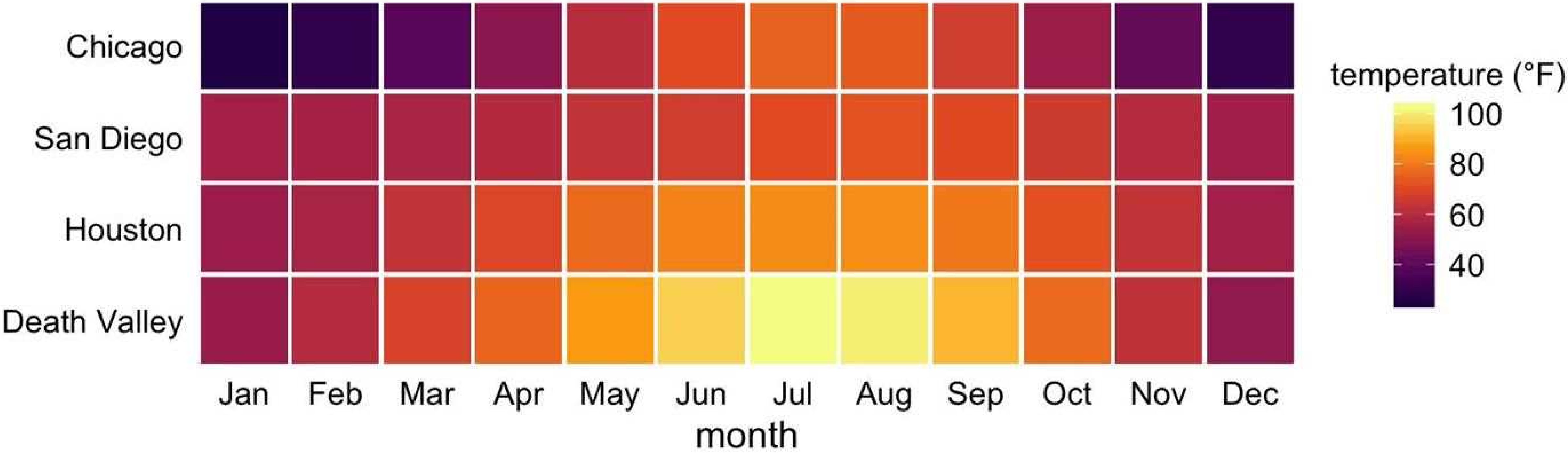


temperature (°F)
100
80
60
40
20
Jan
Apr
Jul
Oct
Jan
month
Death Valley
Houston
San Diego
Chicago

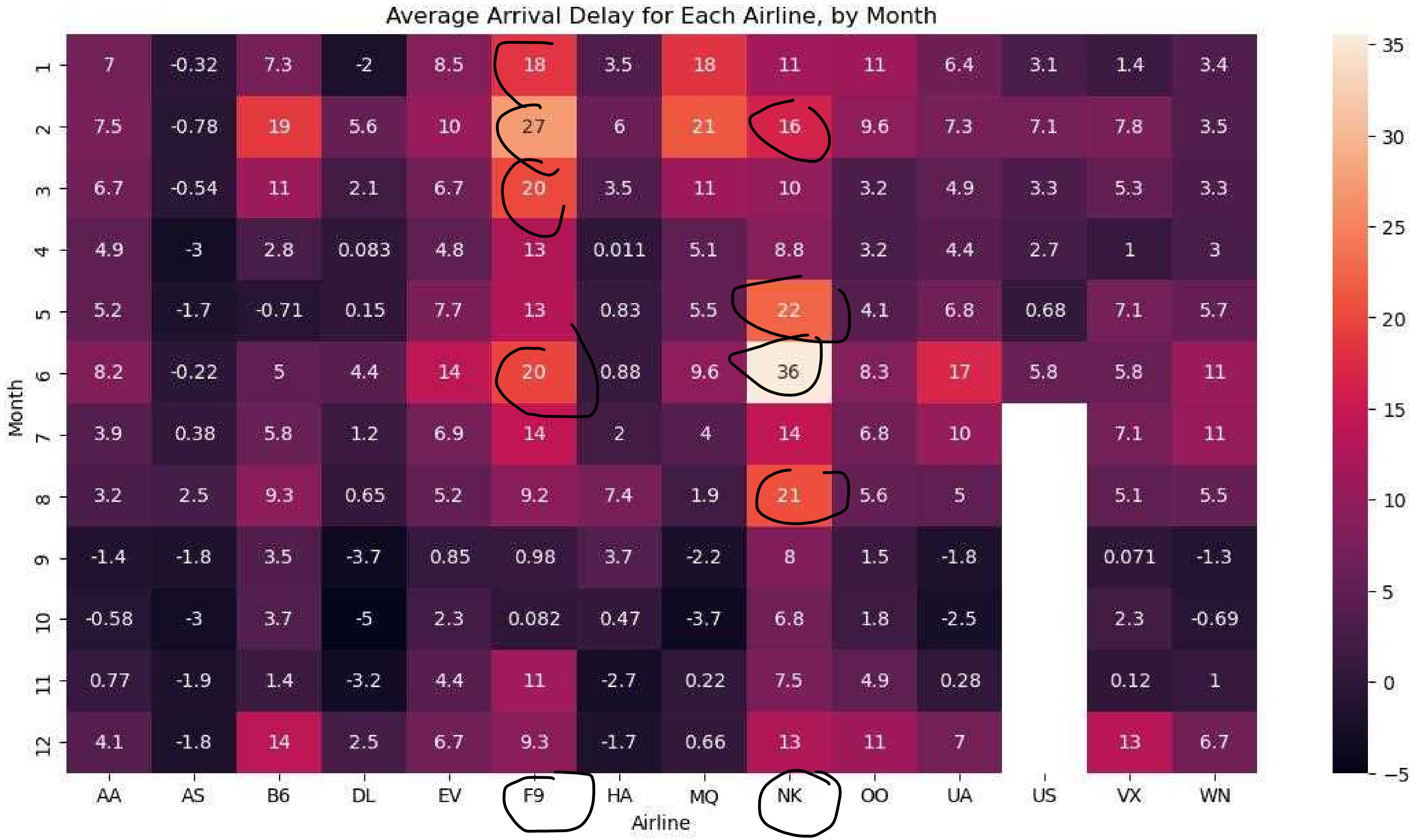
Sales


1400
1200
1000
800
600
400
200
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Product1
Product2
Product3

Heat Map

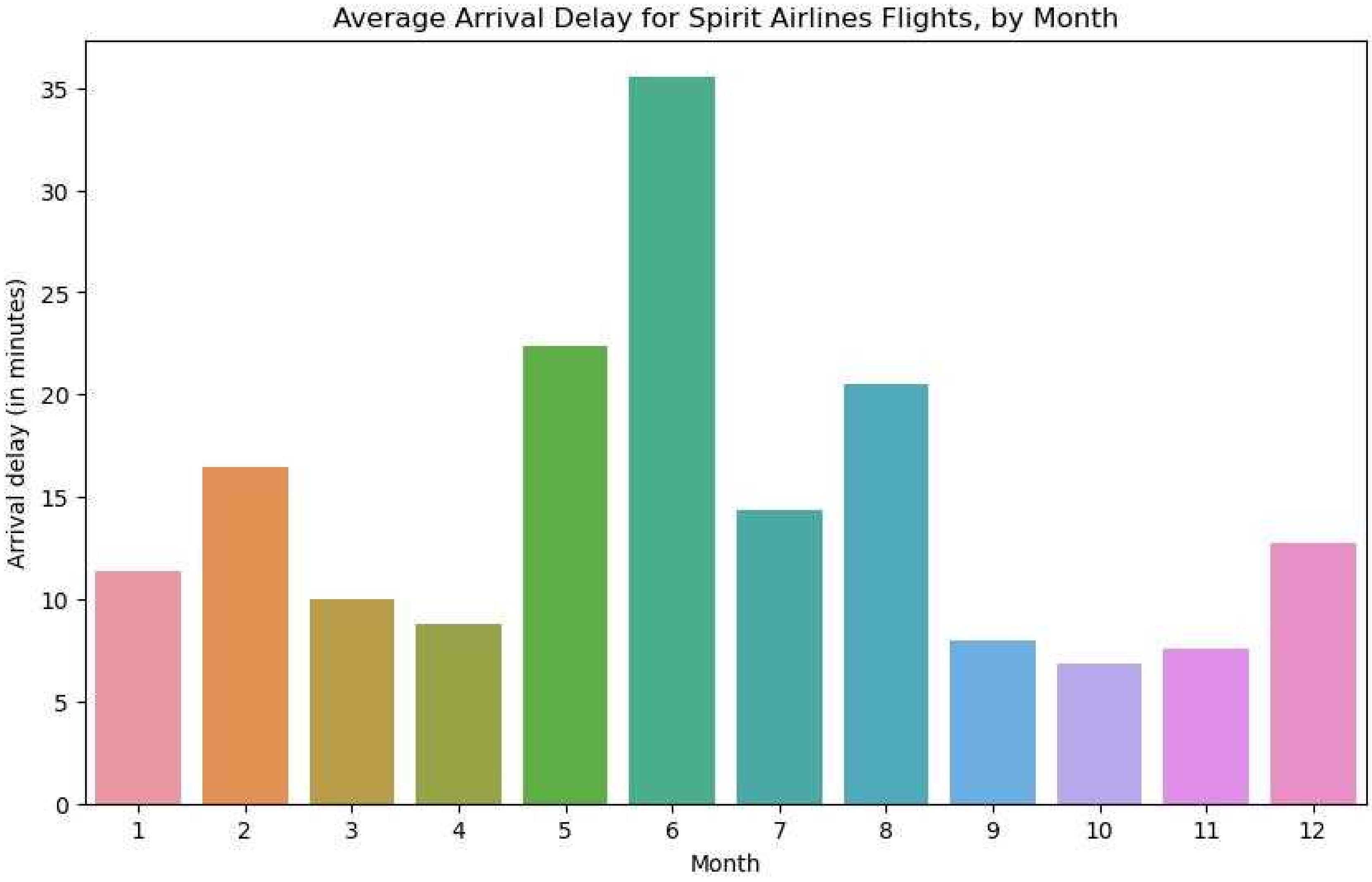

Chicago
San Diego
Houston
Death Valley
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
month
temperature (°F)
100
80
60
40

Arrival time delay Heatmap

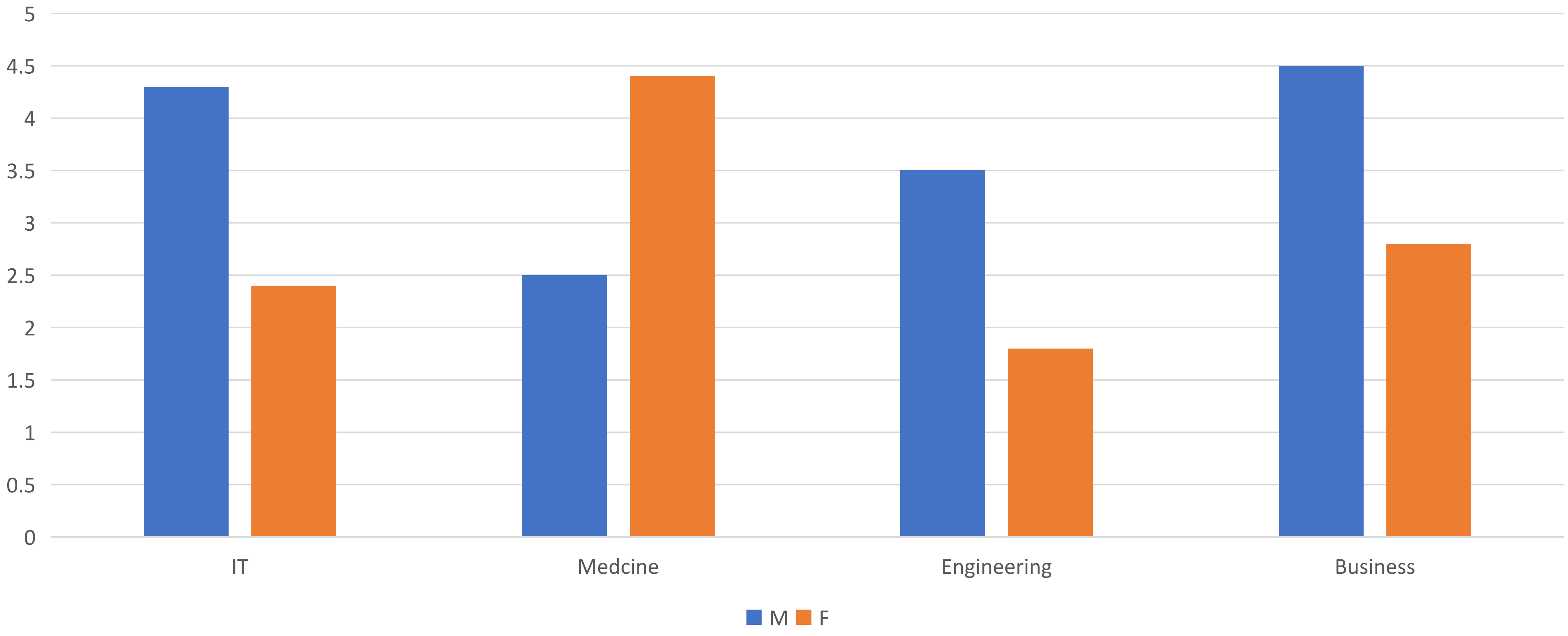


Month	AA	AS	B6	DL	EV	F9	HA	MQ	NK	OO	UA	US	VX	WN
1	7	-0.32	7.3	-2	8.5	18	3.5	18	11	11	6.4	3.1	1.4	3.4
2	7.5	-0.78	19	5.6	10	27	6	21	16	9.6	7.3	7.1	7.8	3.5
3	6.7	-0.54	11	2.1	6.7	20	3.5	11	10	3.2	4.9	3.3	5.3	3.3
4	4.9	-3	2.8	0.083	4.8	13	0.011	5.1	8.8	3.2	4.4	2.7	1	3
5	5.2	-1.7	-0.71	0.15	7.7	13	0.83	5.5	22	4.1	6.8	0.68	7.1	5.7
6	8.2	-0.22	5	4.4	14	20	0.88	9.6	36	8.3	17	5.8	5.8	11
7	3.9	0.38	5.8	1.2	6.9	14	2	4	14	6.8	10		7.1	11
8	3.2	2.5	9.3	0.65	5.2	9.2	7.4	1.9	21	5.6	5		5.1	5.5
9	-1.4	-1.8	3.5	-3.7	0.85	0.98	3.7	-2.2	8	1.5	-1.8		0.071	-1.3
10	-0.58	-3	3.7	-5	2.3	0.082	0.47	-3.7	6.8	1.8	-2.5		2.3	-0.69
11	0.77	-1.9	1.4	-3.2	4.4	11	-2.7	0.22	7.5	4.9	0.28		0.12	1
12	4.1	-1.8	14	2.5	6.7	9.3	-1.7	0.66	13	11	7		13	6.7

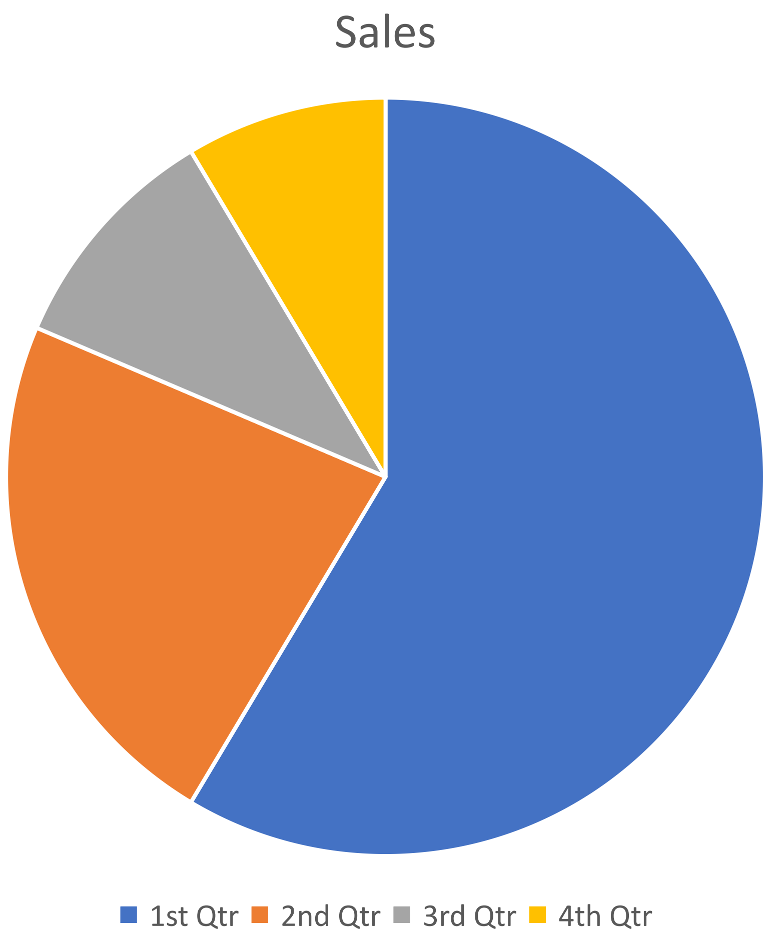
Bar plots

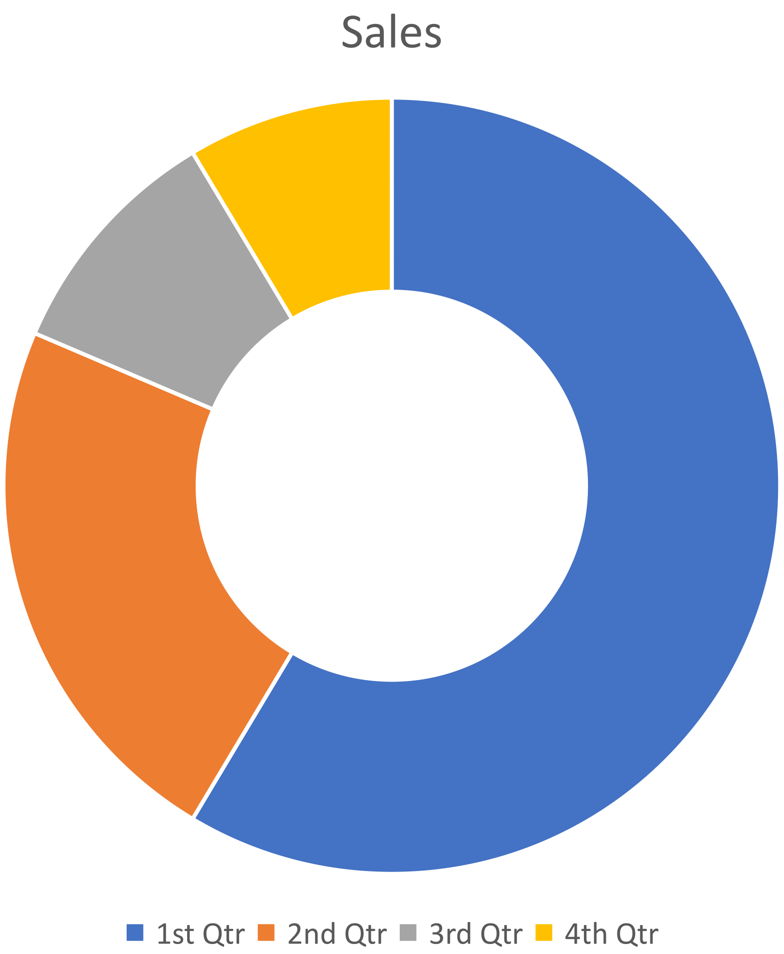

Average Arrival Delay for Spirit Airlines Flights, by Month
Arrival delay (in minutes)
0
5
10
15
20
25
30
35
1
2
3
4
5
6
7
8
9
10
11
12
Month

Bar plots

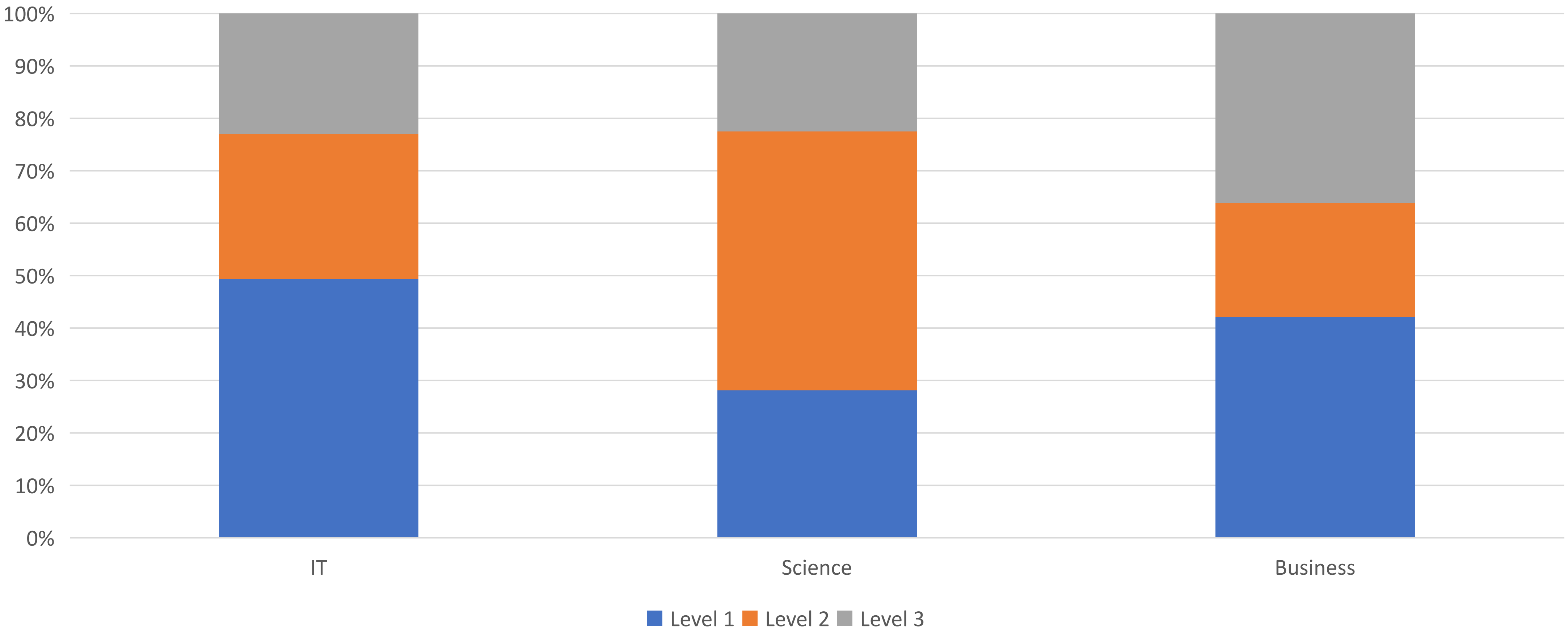


Proportions

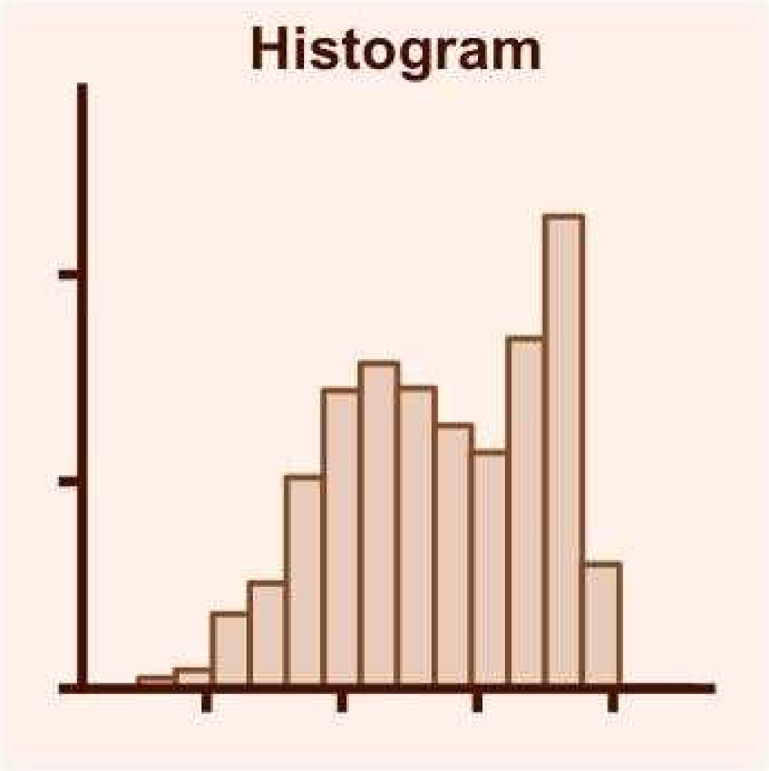


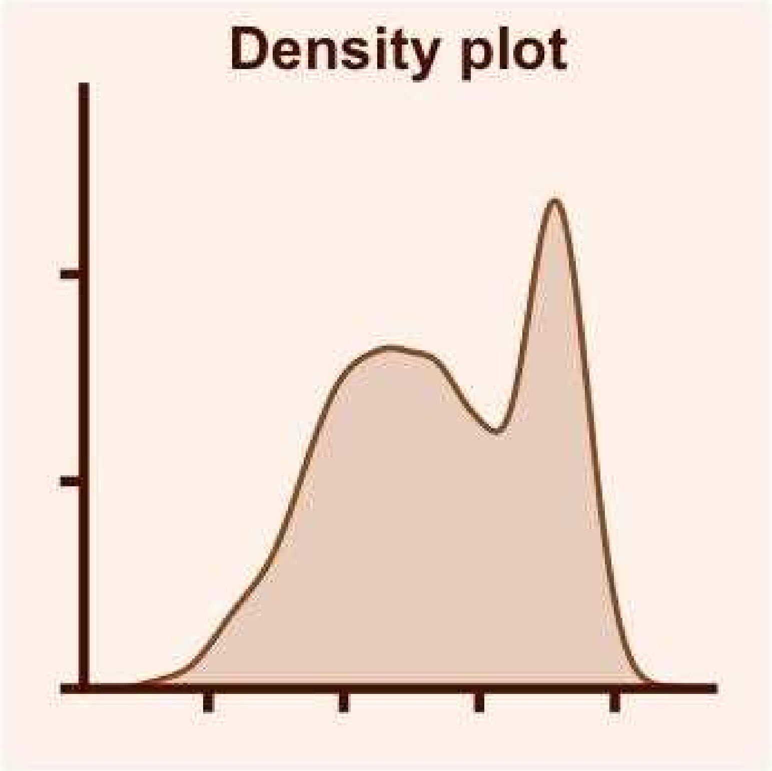


Proportions

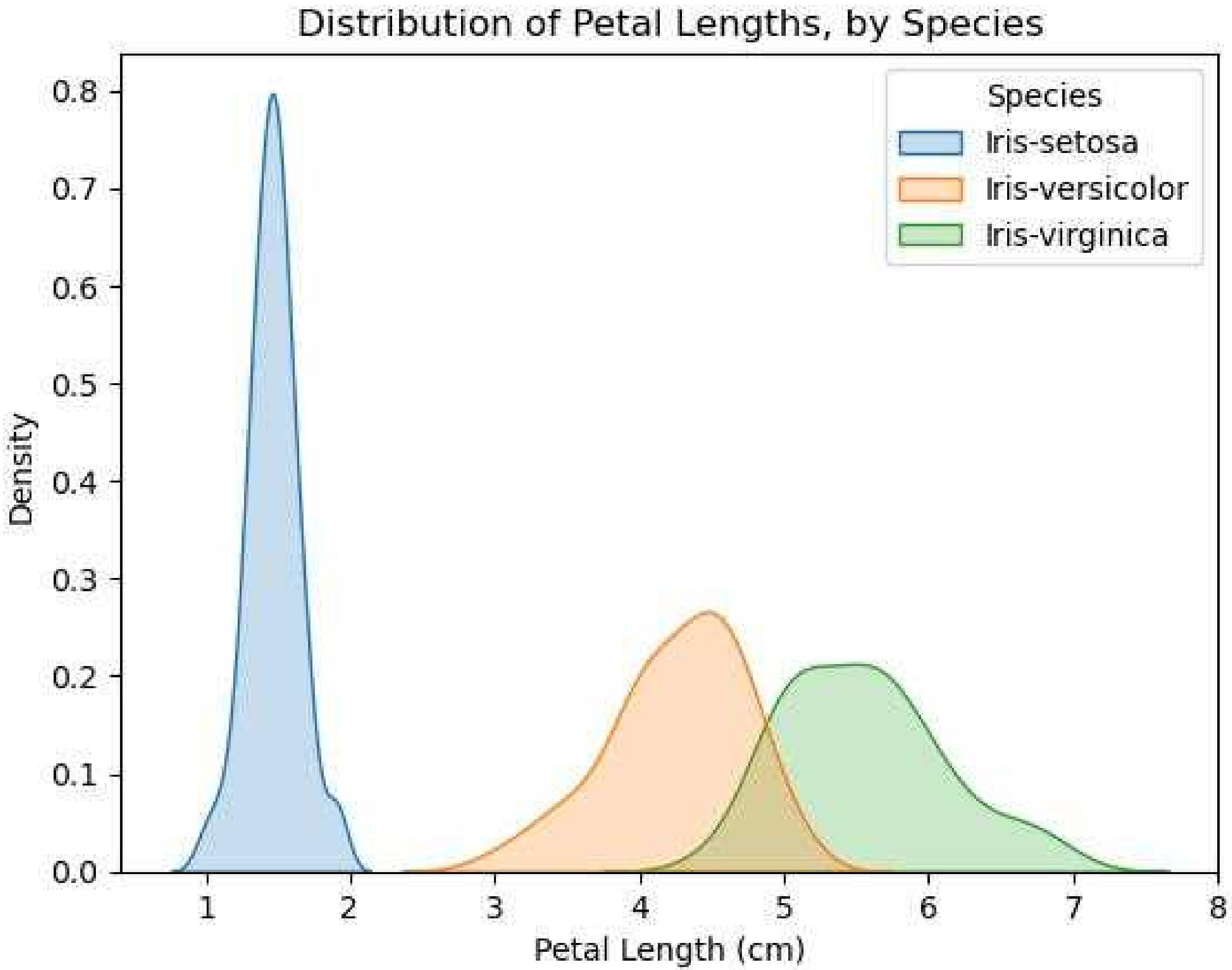


Distributions

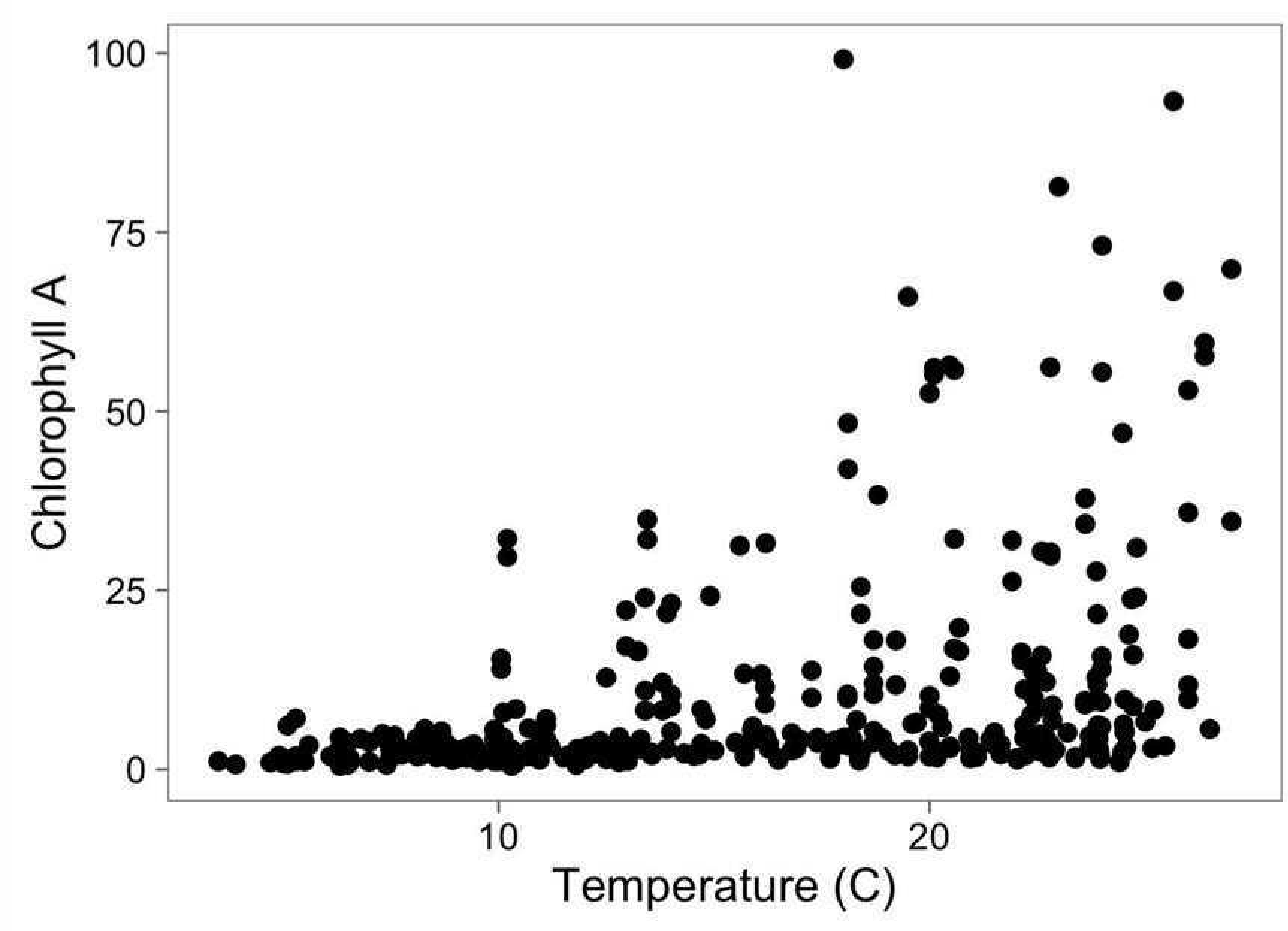




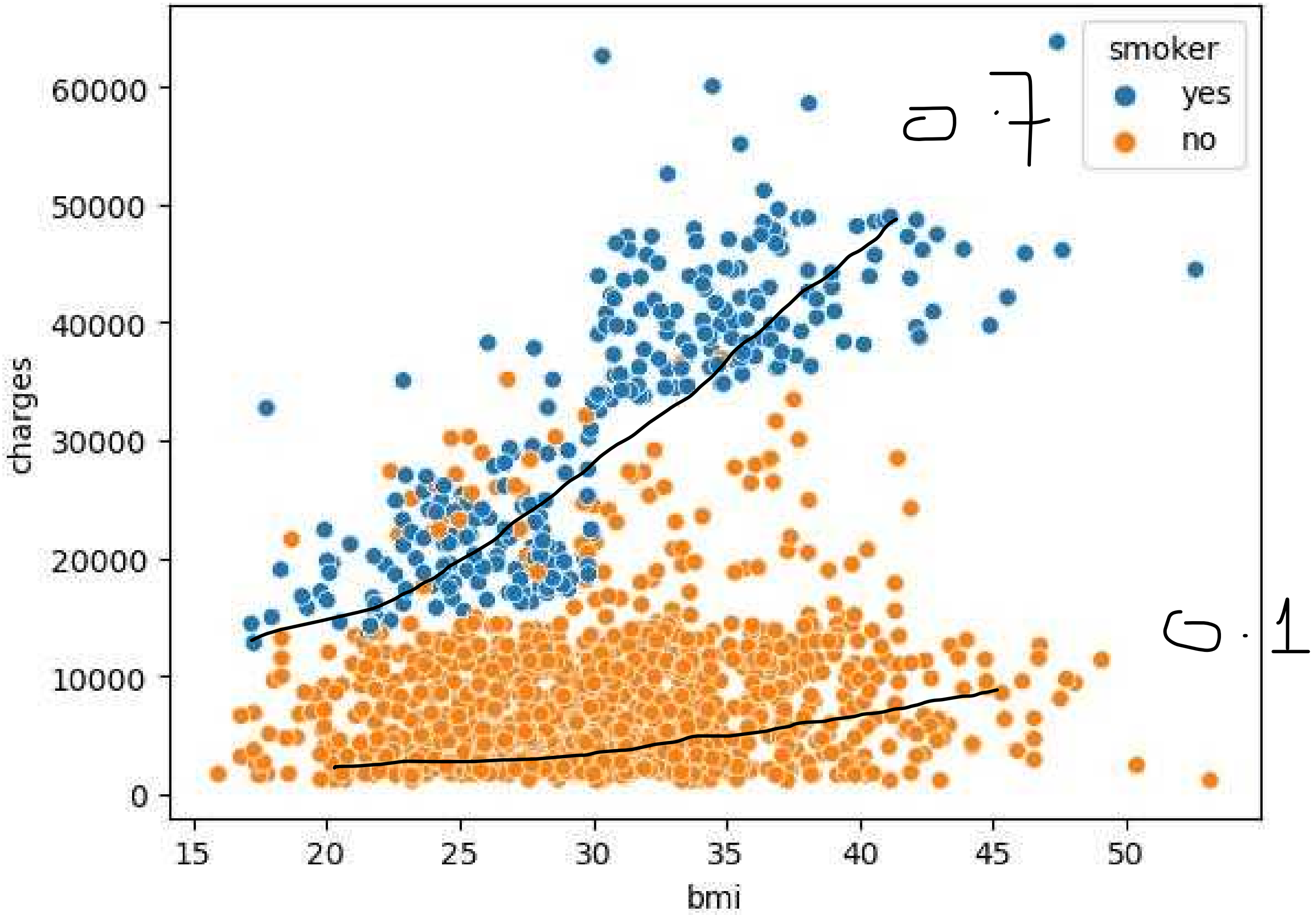
Distribution plot



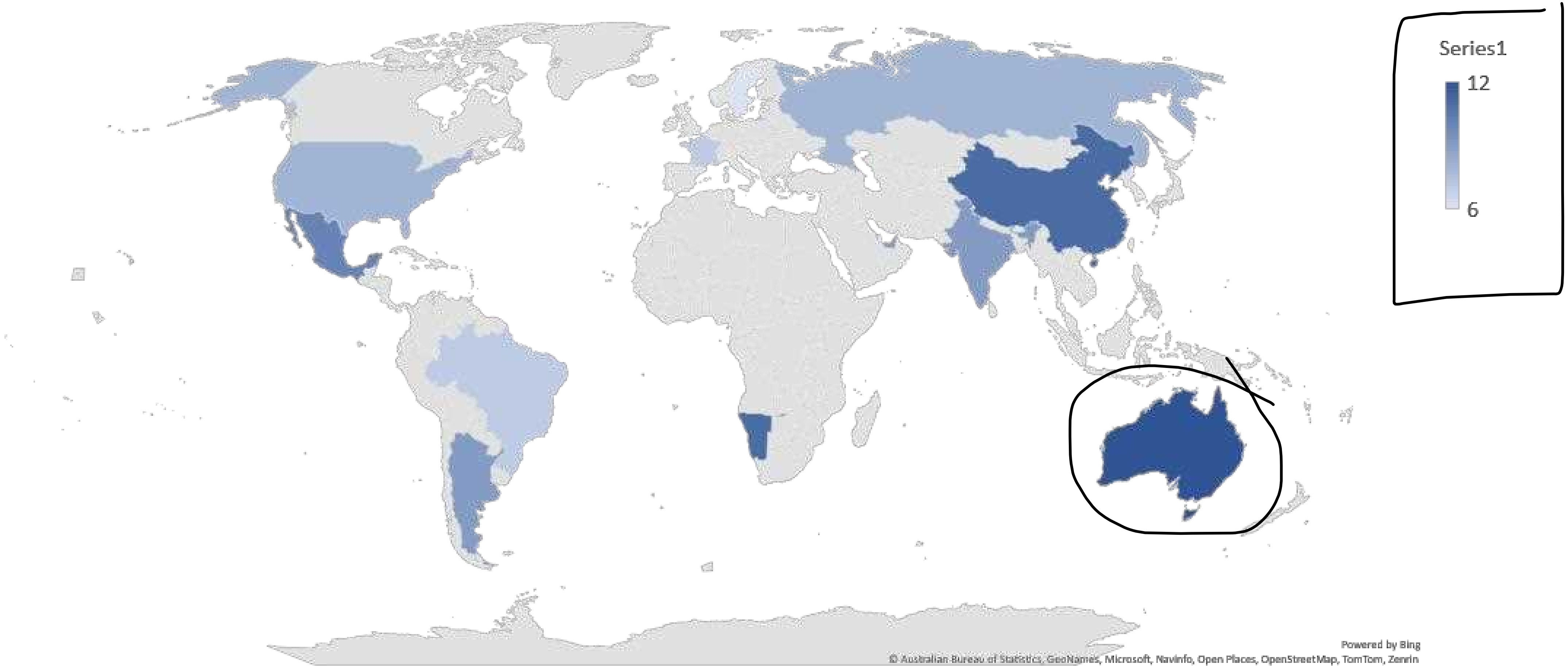
Scatter Plot

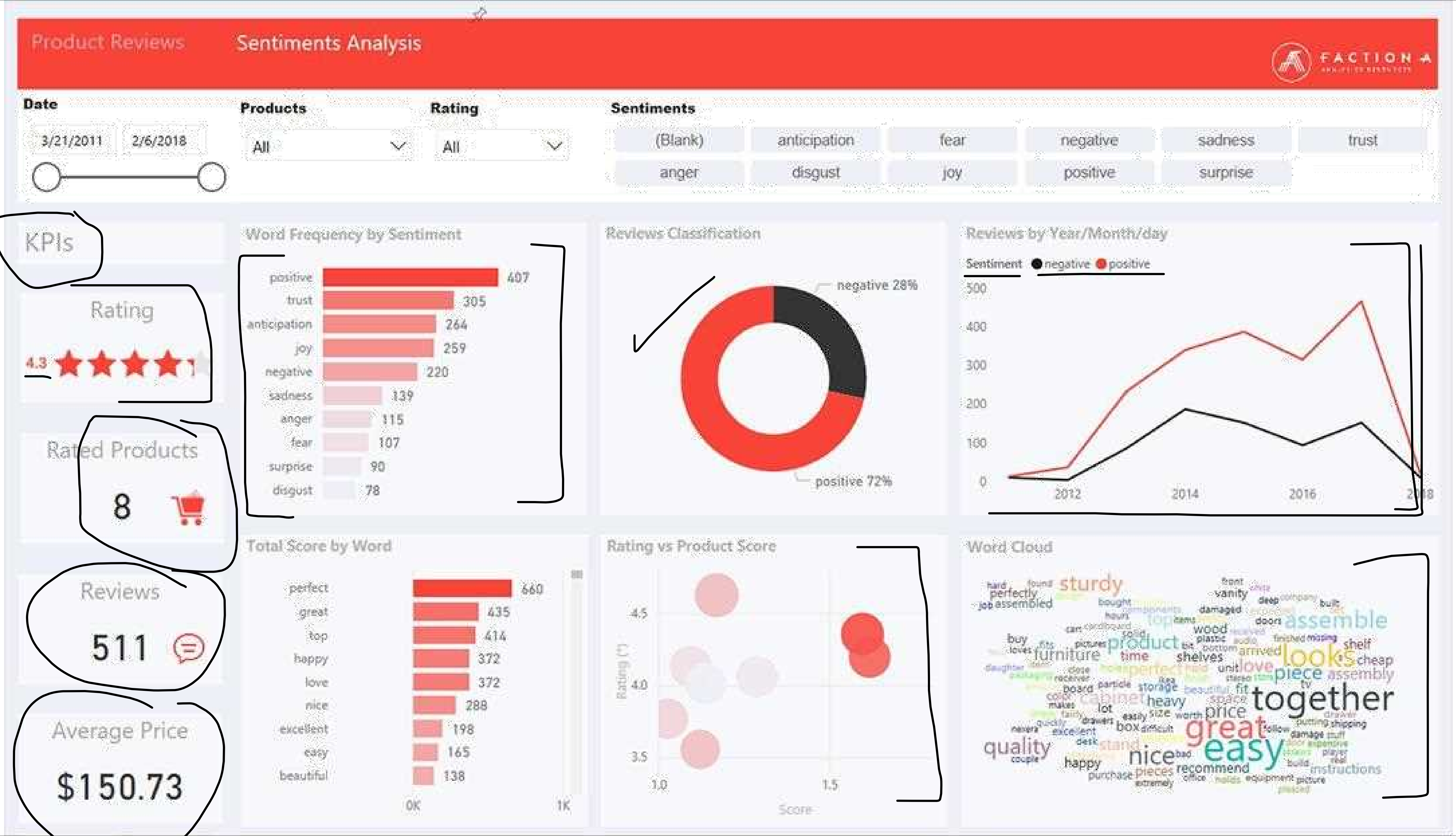


Scatter Plot



Map plot


Series1
12
6
Powered by Bing
© Australian Bureau of Statistics, GeoNames, Microsoft, Navinfo, Open Places, OpenStreetMap, TomTom, Zenrin



Product Reviews

Sentiments Analysis

FACTION A

Date

3/21/2011 2/6/2018

Products

All

Rating

All

Sentiments

(Blank) anticipation fear negative sadness trust

anger disgust joy positive surprise

KPIs

Rating

4.3 ★★★★

Rated Products

8

Reviews

511

Average Price

$150.73

Roadmap for data exploration

1. Organize the data set
2. Find the central point for each attribute
3. Understand the spread of the data for each attribute
4. Visualize the distribution of each attribute
5. Watch out for outliers
6. Understanding the relationship between attributes
7. Visualize the relationship between attributes

Data preparation

Data pre-processing tasks

Main Task	Subtasks	Popular Methods
Data consolidation	Access and collect the data Select and filter the data Integrate and unify the data	SQL queries, software agents, Web services. Domain expertise, SQL queries, statistical tests. SQL queries, domain expertise, ontology-driven data mapping.
Data cleaning	Handle missing values in the data	• Fill in missing values (imputations) with most appropriate values (mean, median, min/max, mode, etc.); • recode the missing values with a constant such as “NA”; • remove the record of the missing value; • do nothing.
	Identify and reduce noise in the data	Identify the outliers in data with simple statistical techniques (such as averages and standard deviations) or with cluster analysis; once identified, either remove the outliers or smooth them by using binning, regression, or simple averages.
	Find and eliminate erroneous data	Identify the erroneous values in data (other than outliers), such as odd values, inconsistent class labels, odd distributions; once identified, use domain expertise to correct the values or remove the records holding the erroneous values.

Data pre-processing tasks

Main Task	Subtasks	Popular Methods
3 Data transformation	Normalize the data	Reduce the range of values in each numerically valued variable to a standard range (e.g., 0 to 1 or −1 to +1) by using a variety of normalization or scaling techniques.
	Discretize or aggregate the data	If needed, convert the numeric variables into discrete representations using range- or frequency-based binning techniques; for categorical variables, reduce the number of values by applying proper concept hierarchies.
	Construct new attributes	Derive new and more informative variables from the existing ones using a wide range of mathematical functions (as simple as addition and multiplication or as complex as a hybrid combination of log transformations).
Data reduction	Reduce number of attributes	Use principal component analysis, independent component analysis, chi-square testing, correlation analysis, and decision tree induction.
	Reduce number of records	Perform random sampling, stratified sampling, expert-knowledge-driven purposeful sampling.
	Balance skewed data	Oversample the less represented or undersample the more represented classes.

Summary

- Important to understand the data available
- Summary statistics provide a good overview
  - Can be deceptive!
- Visualization is a powerful way to understand data
- Data prerpration tasks is important because real-world data is not clean and ordered