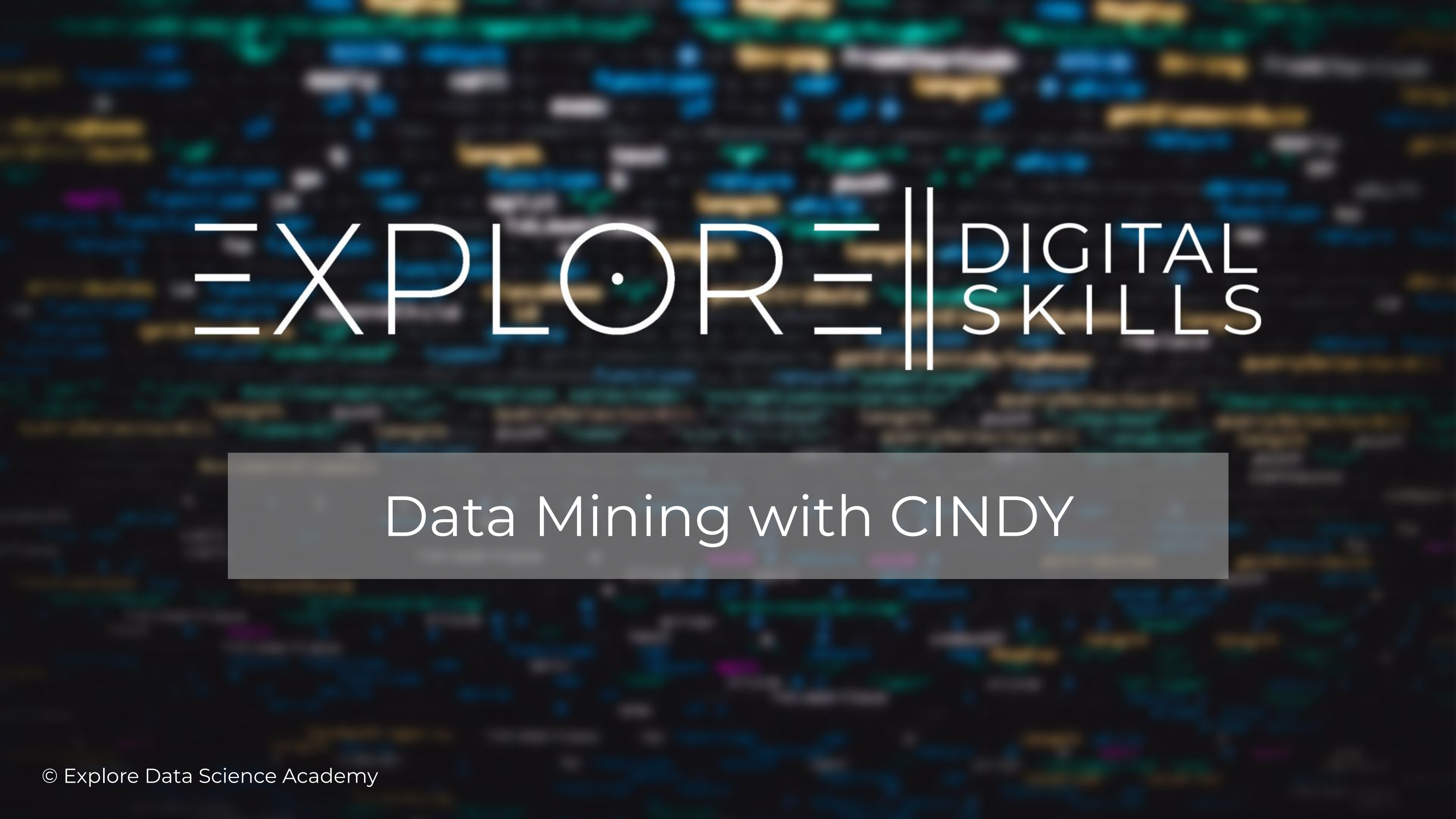


EXPLORE DIGITAL SKILLS

Data Mining with CINDY

Data mining with CINDY

- Introduction to the CINDY framework for optimal visualisations;
- Use the CINDY framework for data visualisation;
- Represent and perform visualisation of different data types in the CINDY framework .

Meet CINDY!

CINDY is a framework that comes with a checklist for understanding the relationship between data.

	Summary Statistics		Relationships between 2 variables				
	Description	Visualise	C	I	N	D	xY
C Categorical	String <=25 uniques	Ordered bar chart 80/20	**Heatmap** **Chi2**	Stacked bar chart	**Boxplot**	Time Series	Categorised Heatmap
I Identifier	String >25 uniques	Rank order SSST		Heatmap Chi2	Rank Order (mean)	Fan (Percentile) Chart	Categorised Heatmap
N Numerical	Integers, Float, Decimal	Histogram Mean, Stdev			**Scatter Plot** Correlation	Time Series (mean)	Graduated Heatmap
D Dates	Timestamp	Time series Stationarity				Histogram Mean diff	Time-lapse Heatmap
x Y Geo-spatial	x, y Lat, lon	Polygons Points					

EXPLORE | DIGITAL SKILLS

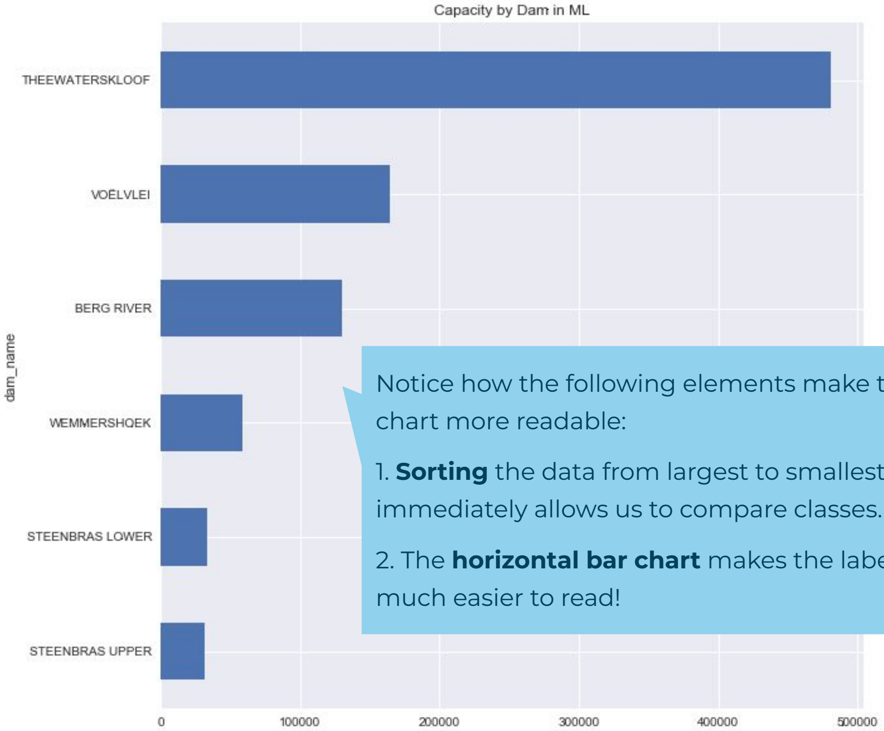
CINDY Checklist

CINDY comes with a checklist for understanding the relationships between data.

	Summary Statistics	Relationships between 2 variables	
		Categorical	Numerical
C Categorical	**Barplot** □ Nulls / missing data □ Unique values □ Lookups / Identifiers □ Links to other datasets	**Chi2-test / Heatmap** □ Indicator variables □ Equal buckets	**Boxplot** □ Boxplots (for >10) □ Outliers □ Ordering is important
I Identifier			
N Numerical	**Histogram (KDE)** □ Nulls / missing data □ Mean & Standard deviation □ Outliers	**Time Series** □ Nulls / missing data □ Data Formats	**Scatter** □ Linear correlation □ Beta coefficient □ Axis □ Non-linearity
D Dates			
X Y Geo-spatial			

Categorical Variables – Bar Chart

Barchart for Categorical Variables



Notice how the following elements make this chart more readable:

1. **Sorting** the data from largest to smallest immediately allows us to compare classes.

2. The **horizontal bar chart** makes the labels much easier to read!

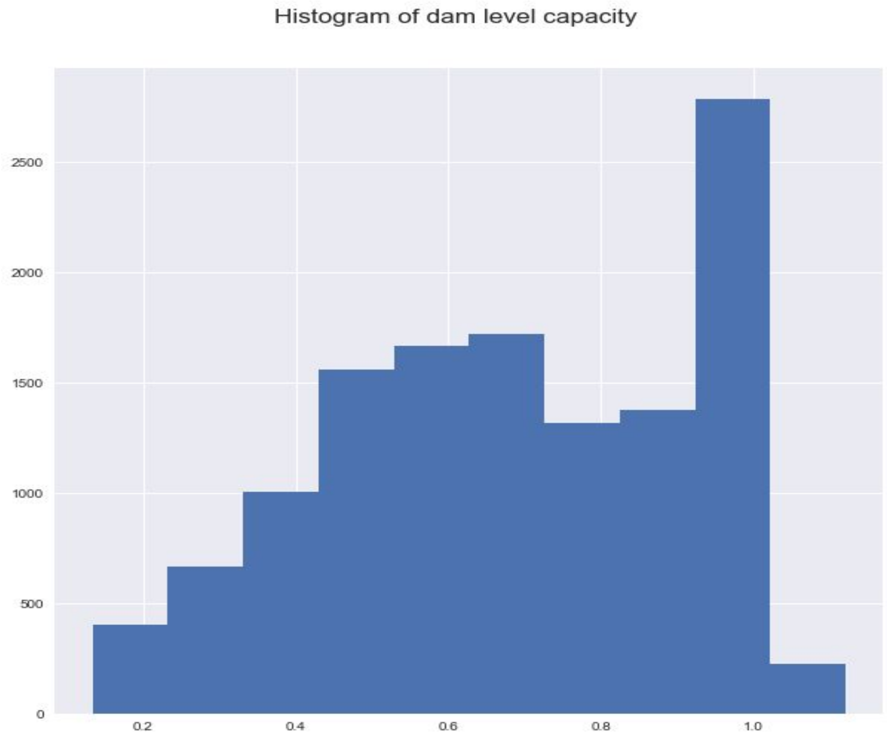
```
pandas.DataFrame.groupby('cat').count().plot()
```

Things to look out for in bar charts:

- Count **unique values**.
- Check for **nulls**.
- Apply **80/20 principle** on categorical variables- Use this to focus analysis on the most important categories.
- Look for **groupings/lookups** – combine categorical variables into more interpretable combinations and results.
- Categorical variables provide a good way to **link data between datasets**.

Numerical Variables – Histogram

Histogram for Numerical Variables



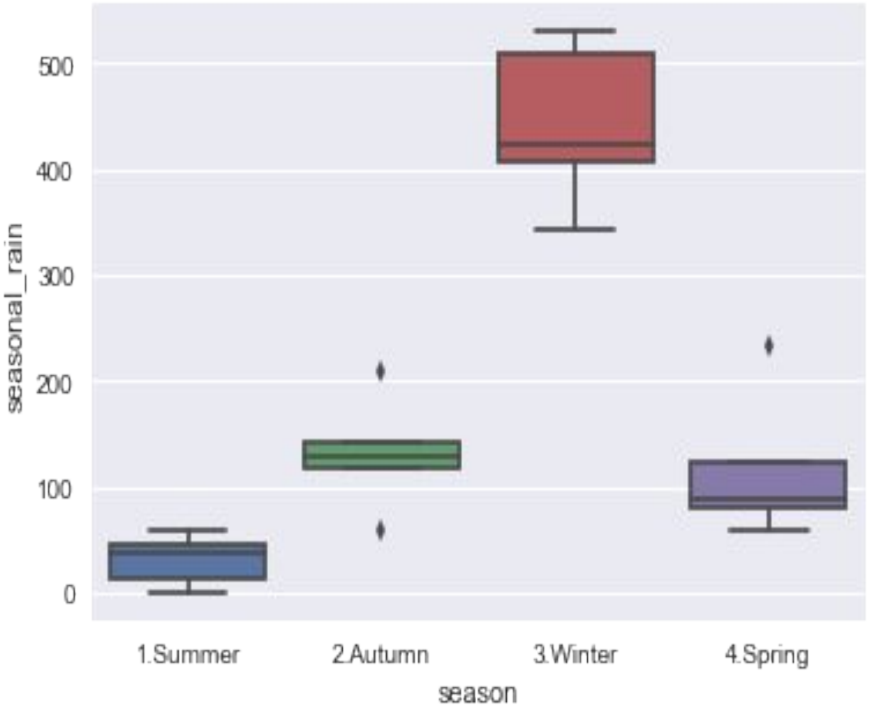
```
pandas.DataFrame.hist()
```

Things to look out for in histograms:

- Check for **nulls.**
- **Summary statistics** are very helpful to understand numerical variables:
  - **Mean** and **standard deviation.**
  - Percentiles (especially the **median**).
- Identify the closest **distribution function.**
- **Outliers identification.**

Numerical Variables by Category - Box Plot

Box plots



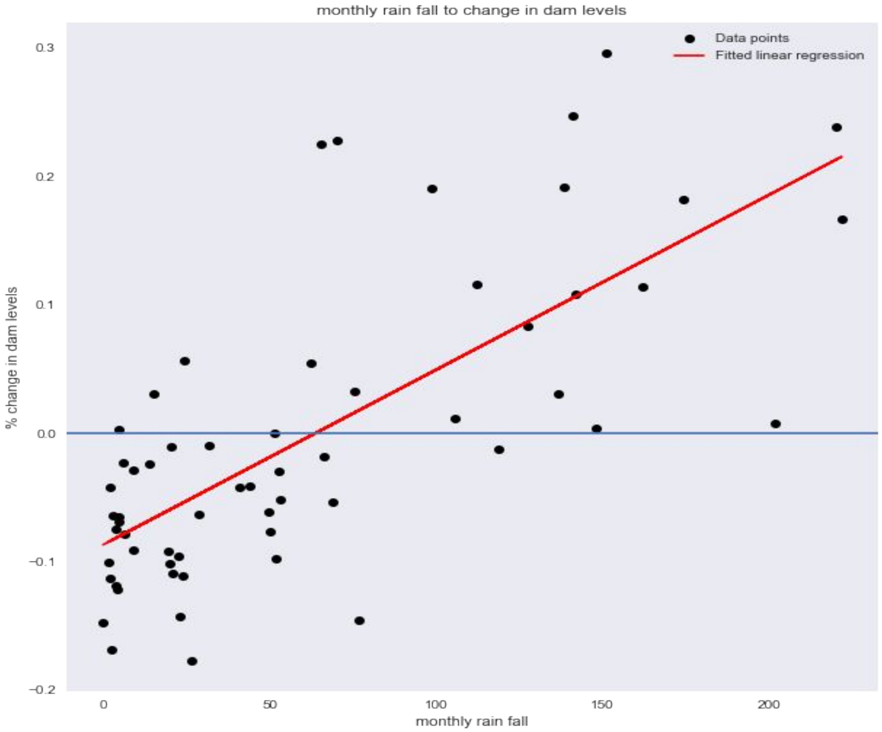
```
seaborn.boxplot()
```

Things to look out for in a box plot:

- Box plots provide information about the **5 number summary** of a dataset:
  - minimum value
  - first quartile (Q1)
  - median
  - third quartile (Q3)
  - maximum value
- Often used for descriptive analyses or during the preliminary investigation of a large data set.
- **Box plots** are used to indicate whether the distribution in a dataset is skewed or used for the identification of outliers in the dataset.

Relationships between Numerical Variables - Scatter Plot

Scatter Plot



Things to look out for in scatterplots:

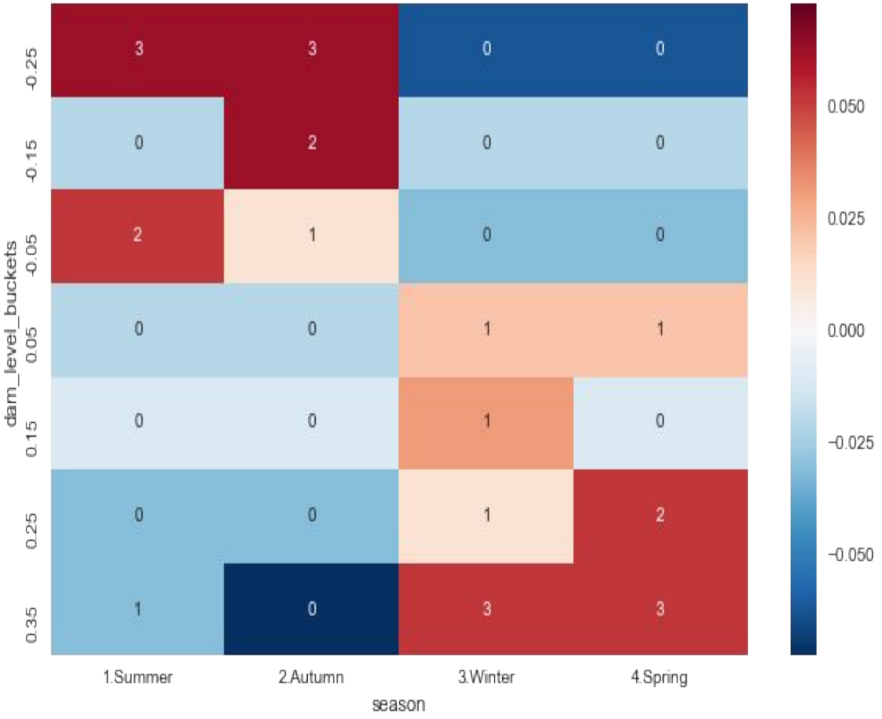
- A **line of best fit** is used to assess the **relationship of variables** in the dataset. The line of best fit (**linear regression)** equation is given as:

$$y = \alpha + \beta x$$

  - **β -** impact of independent variable (x) on the dependent variable (y); this will indicate the slope of the line of best fit.
  - **α -** indicates the y intercept (when x=0).
  - R^2 - the **coefficient of determination**. This indicates the percentage of variation explained by the other variable
  - Outliers directly impact results of linear regression.

Relationships between Numerical Variables - Contingency Table with a Heatmap

Contingency Tables with Heatmap



dam_level_buckets \ season	1.Summer	2.Autumn	3.Winter	4.Spring
-0.25	3	3	0	0
-0.15	0	2	0	0
-0.05	2	1	0	0
0.05	0	0	1	1
0.15	0	0	1	0
0.25	0	0	1	2
0.35	1	0	3	3

Things to look out for in contingency tables:

- **Contingency table** tabulates the state of a combination of 2 or more categorical variables.
- **Chi^2 test (test for independence)** helps determine if the **distribution** of one **categorical variable** matches another or differs from another and is calculated using the equation:

$$chi^2 = \sum \frac{(Observed\ - expected)^2}{expected}$$

- **Heatmap** - cells are shaded according to the difference in the observation vs expectation counts.
- In the example to the left, **Red cells** represent combinations based on a higher probability of occurrence.

Conclusion

In this train you have learned how to:

- Use the CINDY framework to represent and analyse your data.
- Integrate the use of the CINDY framework to aid in selection of the best method for visualisation and representation of data.

Appendix

Additional sources:

- Data mining
- Descriptive statistics
- Linear Regression