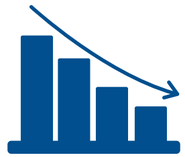
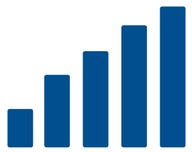




Using Better Visualisations to Elevate Your Data Insights

Mansi Aggarwal | WWCode Dev Summit: BlockDataPy | 2023

Hi, I am Mansi!

2022-23 Fellow @ WWCode Data Science

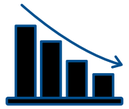
Junior Data Scientist @ Longview, Melbourne

Masters student at Monash University

Interested in bioinformatics and started working on my master's thesis project in melanoma research this semester

Avid sitcom enthusiast

AGENDA



Why visualisation?

What are the pillars of good visual storytelling?

How to improve a standard chart?

How do we derive deeper insights using visualisations?

Why visualisation?

Enhanced comprehension

Detect anomalies and outliers

Uncover patterns and trends

Improved decision-making

Effective communication

Support storytelling

Four pillars of good visual storytelling

PURPOSE

- Why am I creating this visualization?
- Who is it for?
- What do they need to understand?
- What actions do I need to enable?
- What is the most important take-away message?

CONTENT

PURPOSE

- Informed by purpose!
  What data matters?
- What's excluded is as important as what's included.

- Bars for comparison
- Line charts for time, continuity
- Stacked bar charts, pie charts for composition
- Scatter plots for correlation
- Heatmaps for volume of locations/events

FORMATTING

STRUCTURE

CONTENT

PURPOSE

- Helps highlight what's important
- Adds appeal and focus
- Just like icing on a cake!

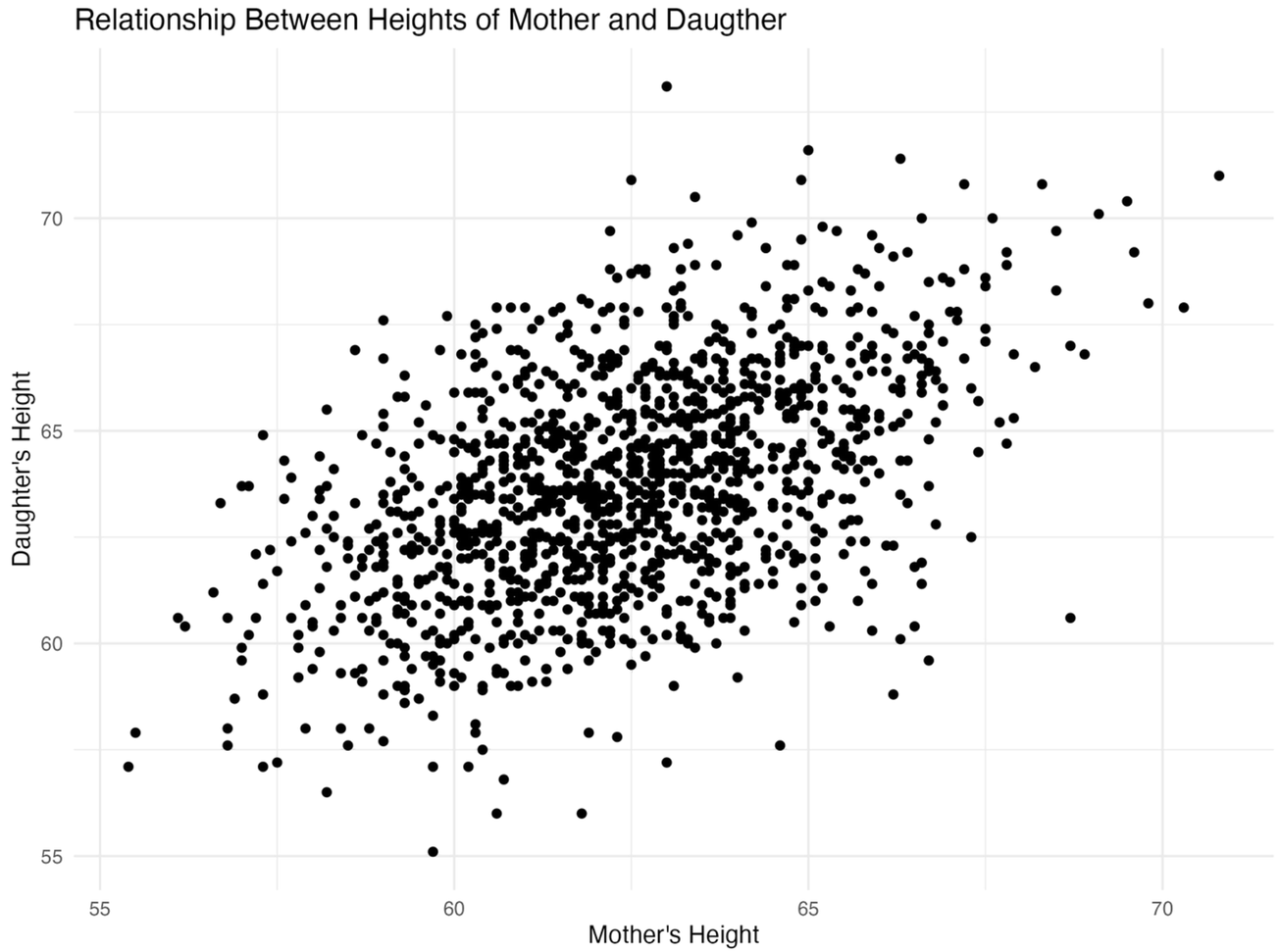
Raw data hardly conveys useful information

Mh <dbl>	Dh <dbl>
59.7	55.1
58.2	56.5
60.6	56.0
60.7	56.8
61.8	56.0
55.5	57.9
55.4	57.1
56.8	57.6
57.5	57.2
57.3	57.1

Good visualisations make information accessible

We now know that

- this data has two variables - mothers' heights and daughters' heights
- there is a strong positive relationship between the two

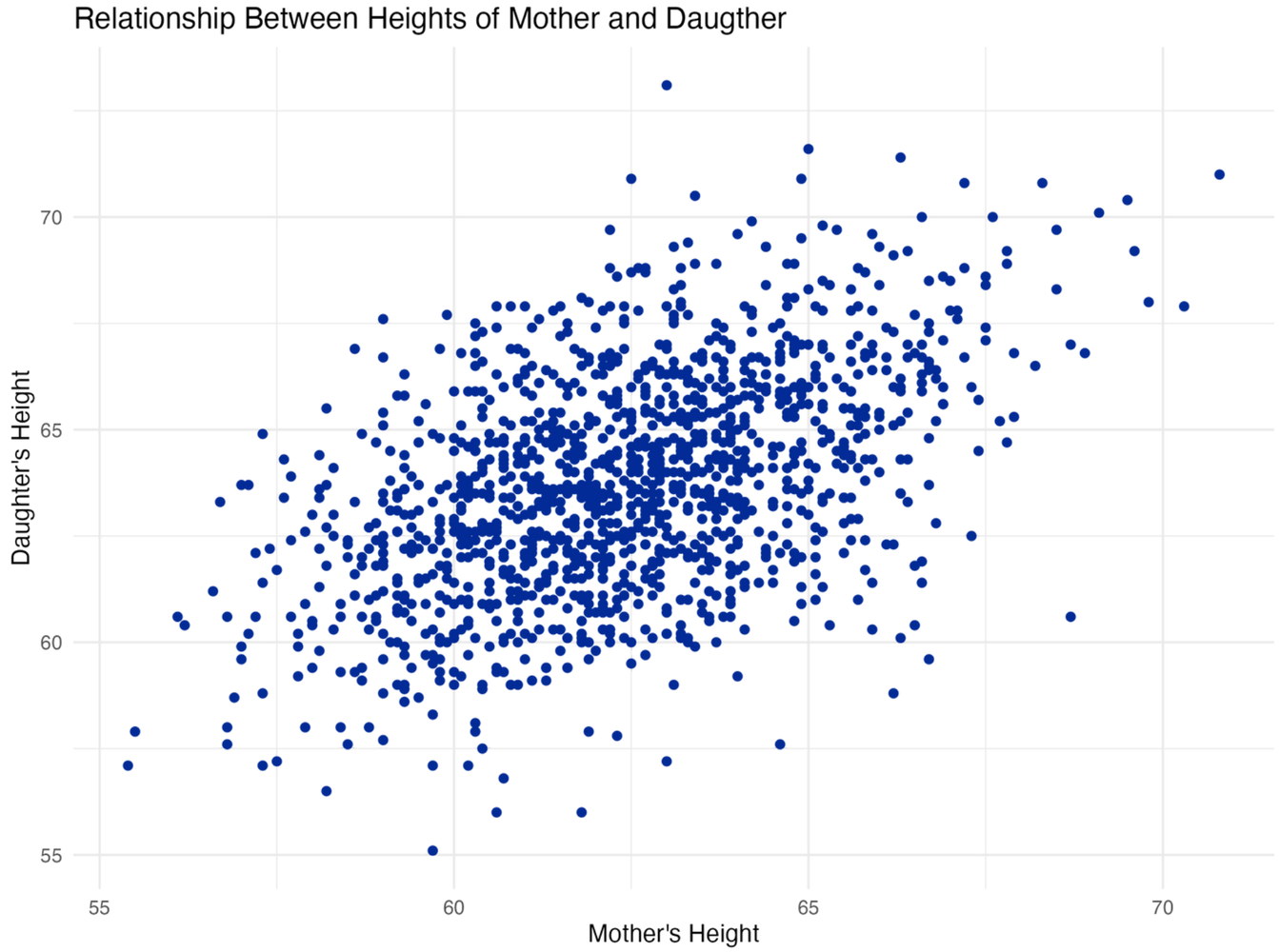


But how to improve a standard chart?

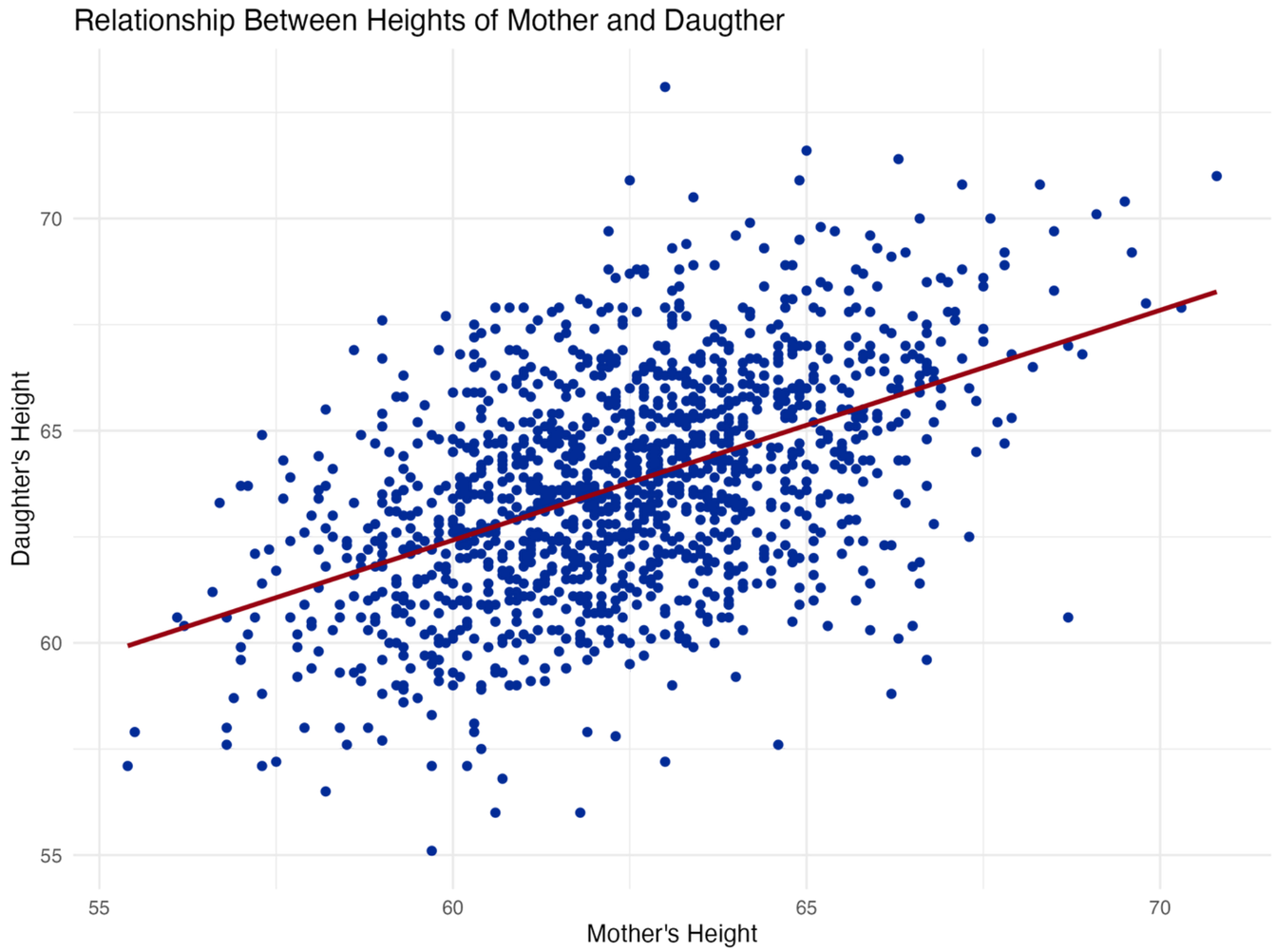
But how to improve a standard chart?

- Less is more - start simple!
- Use clear labels, titles (and subtitles)
- Customise your charts - using non-default colours and settings
- Emphasise key information
- Test for accessibility
- Seek feedback

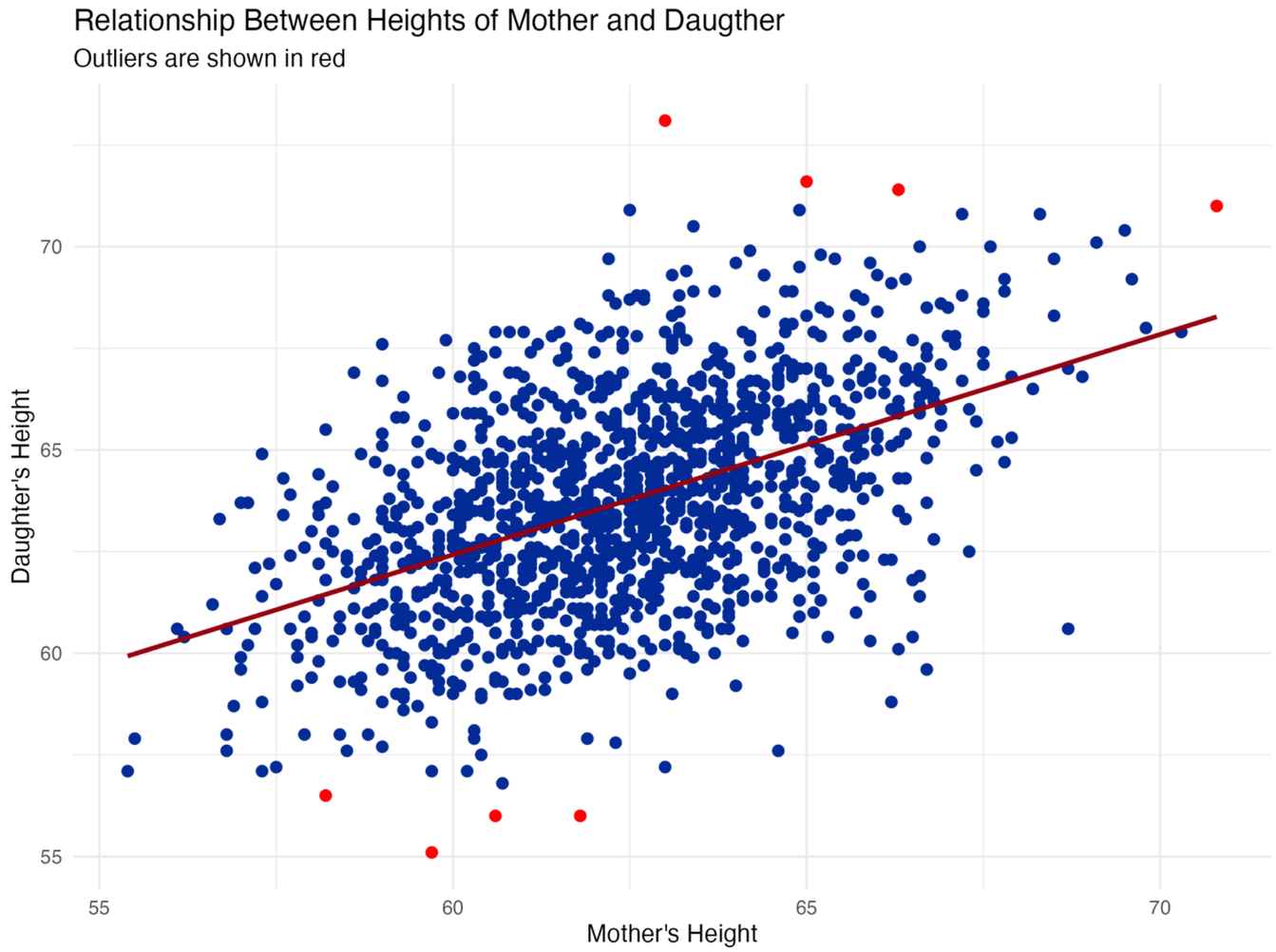
- Non-default colour



- Non-default colour
- Emphasising key information - a strong positive relationship between our two variable

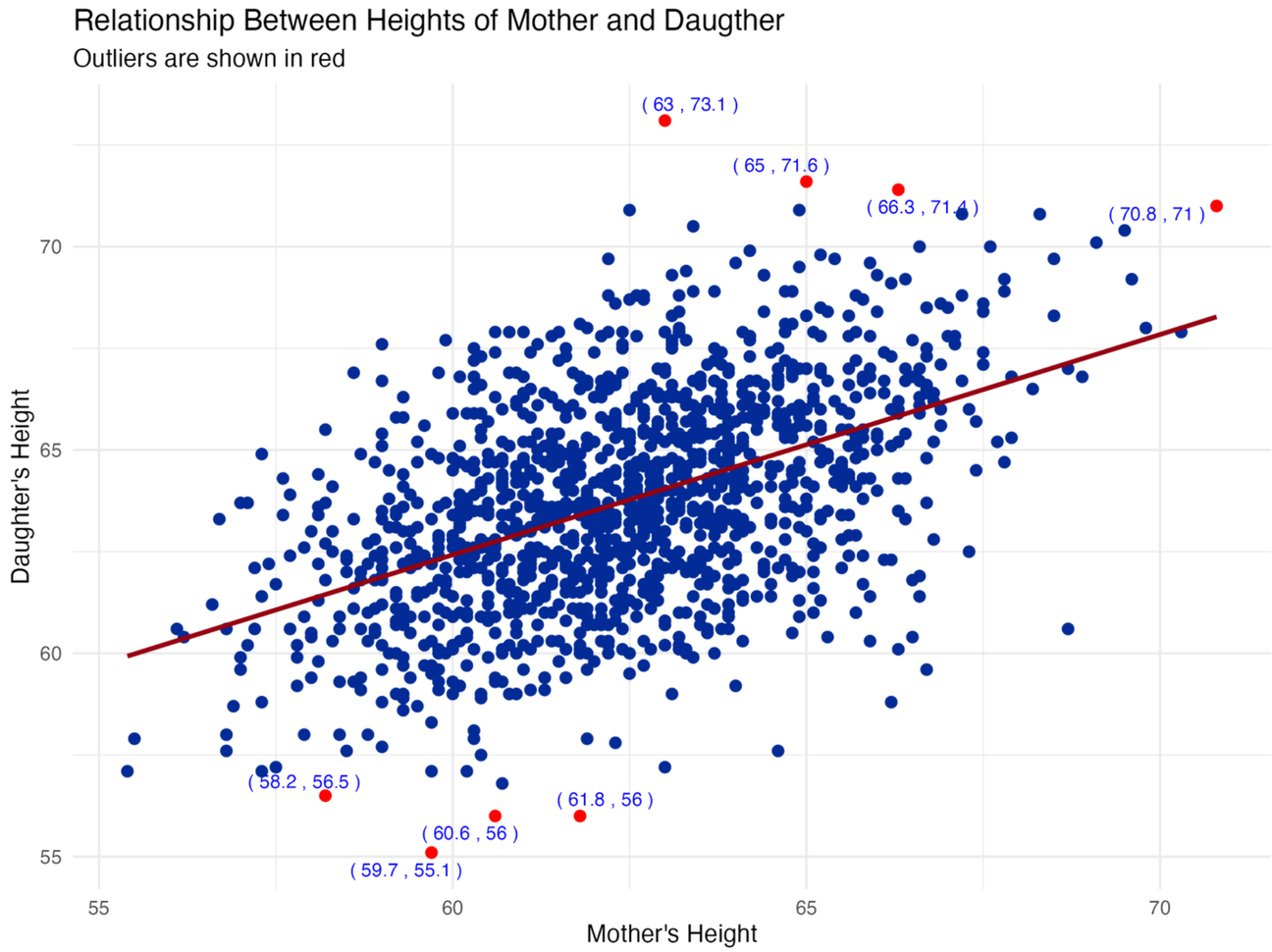


- Non-default colour
- Emphasising key information - a strong positive relationship between our two variable
- Sprucing it up - showing outliers in a different colour!



Start small and work your way up, and always have a backup plan!

If you have more time - tinker with your baseline charts. Here, I annotated the outliers with the x and y coordinates



How do we derive deeper insights using visualisations?

We are going to look at an example from the world of bioinformatics!

You do not have to be an expert to follow this - just see this from a lens of how different visualisations can offer meaningful insights which might otherwise be hard to decipher from large tables!

What exactly will we be looking at?

My report aimed to provide a thorough review of the 2020 Nature article by Bojkova et al. titled "SARS- CoV-2 Infected Host Cell Proteomics Reveal Potential Therapy Targets" and understanding the molecular interactions between the virus and host cells is crucial for developing effective therapeutic strategies.

Things to remember as we move forward?

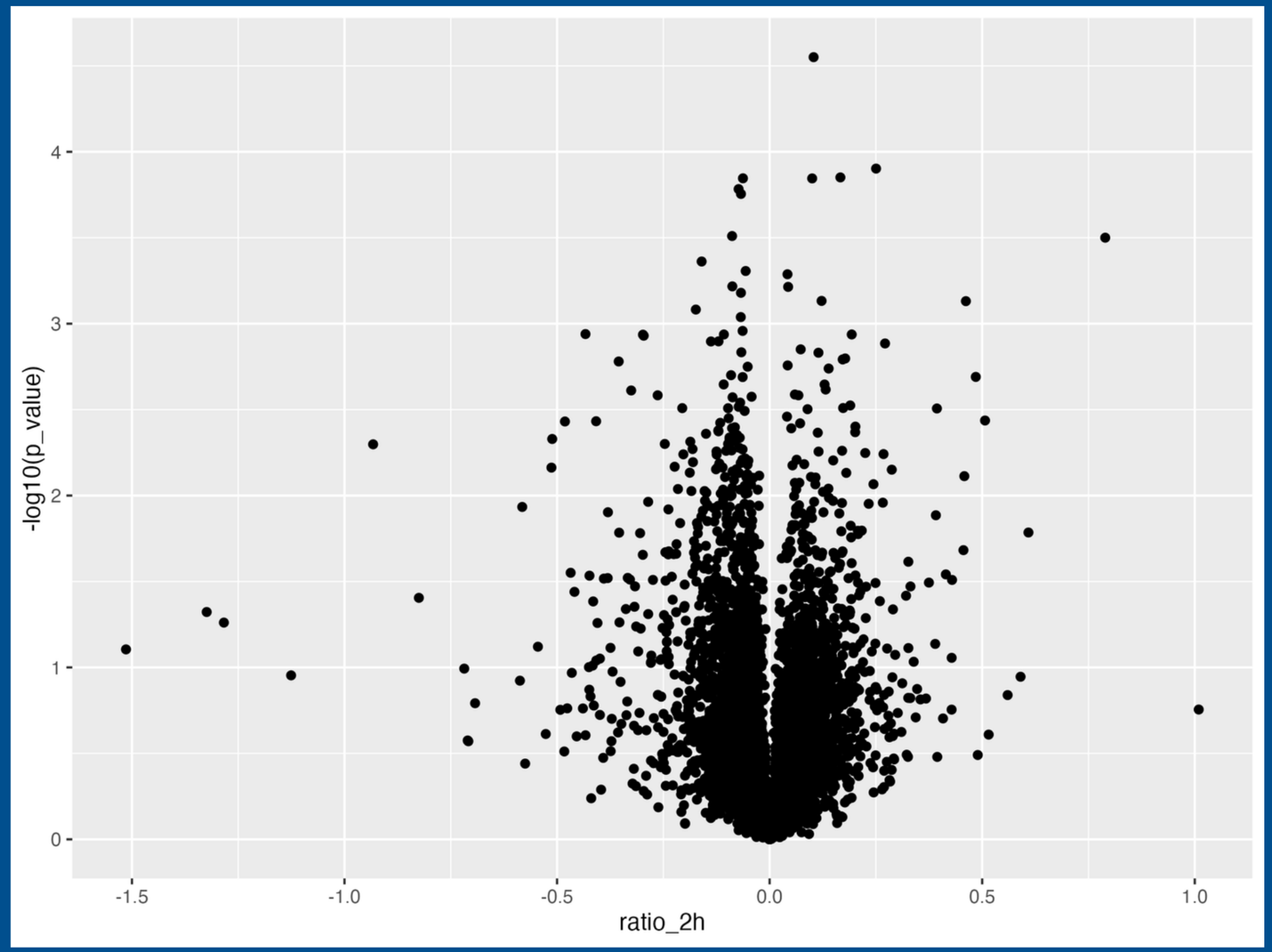
Don't focus on understanding the meaning of the biological terms being used - take everything at face value!

For instance, upregulation means "increase in the state of something" and downregulation means "decrease in the state". "Significant" means "p-value less than 0.05"!

Our dataset

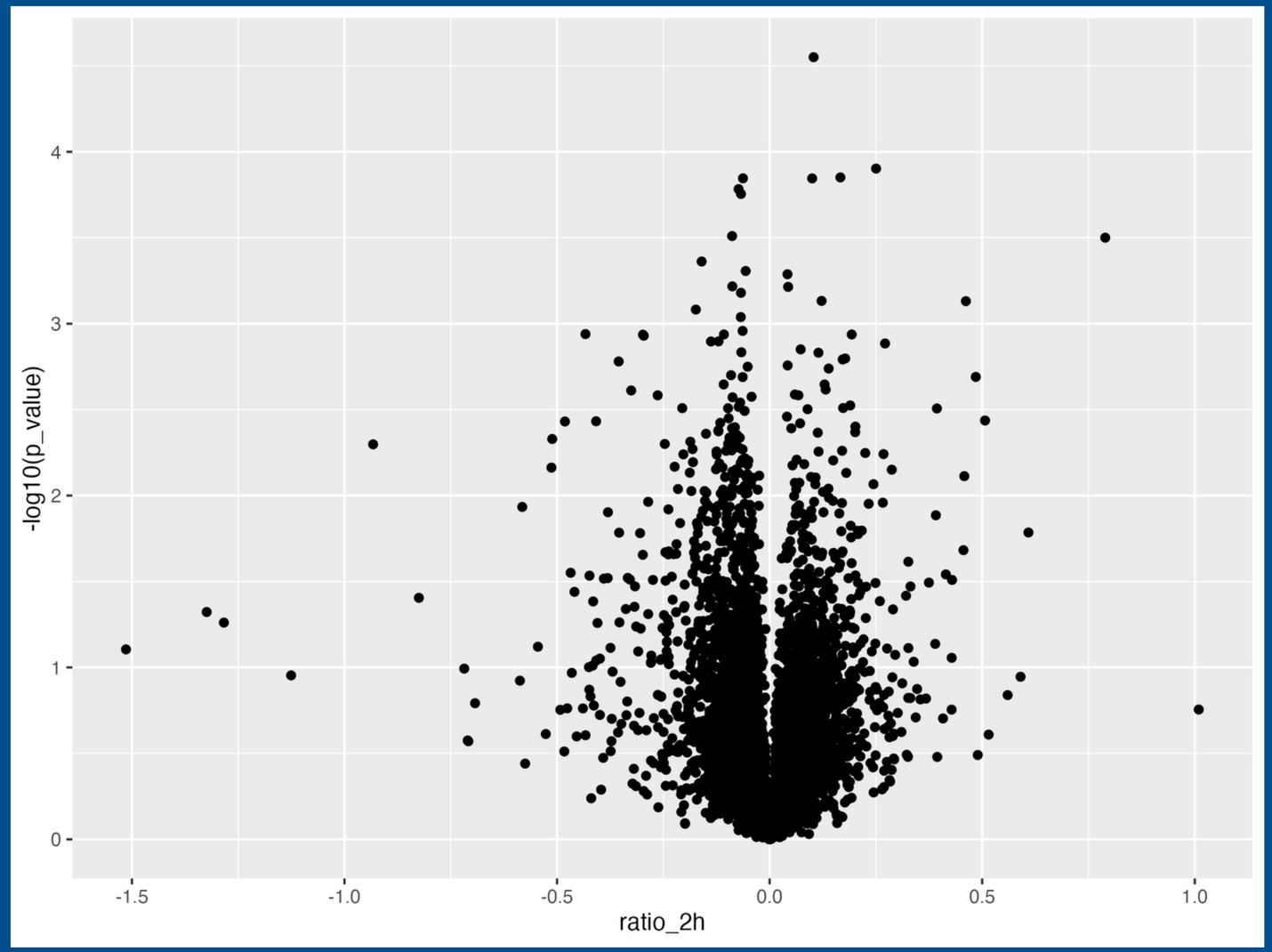
We monitored host cell response to infection at various periods – **2, 6, 10, and 24 hours** – following infection and studied **2,715** significant proteins.

This was my starting point - I plotted all 2715 proteins with their p-value on the y-axis and ratio at the 2-hour time point on the x-axis!



This was my starting point - I plotted all 2715 proteins with their p-value on the y-axis and ratio at the 2-hour time point on the x-axis!

What is going on?



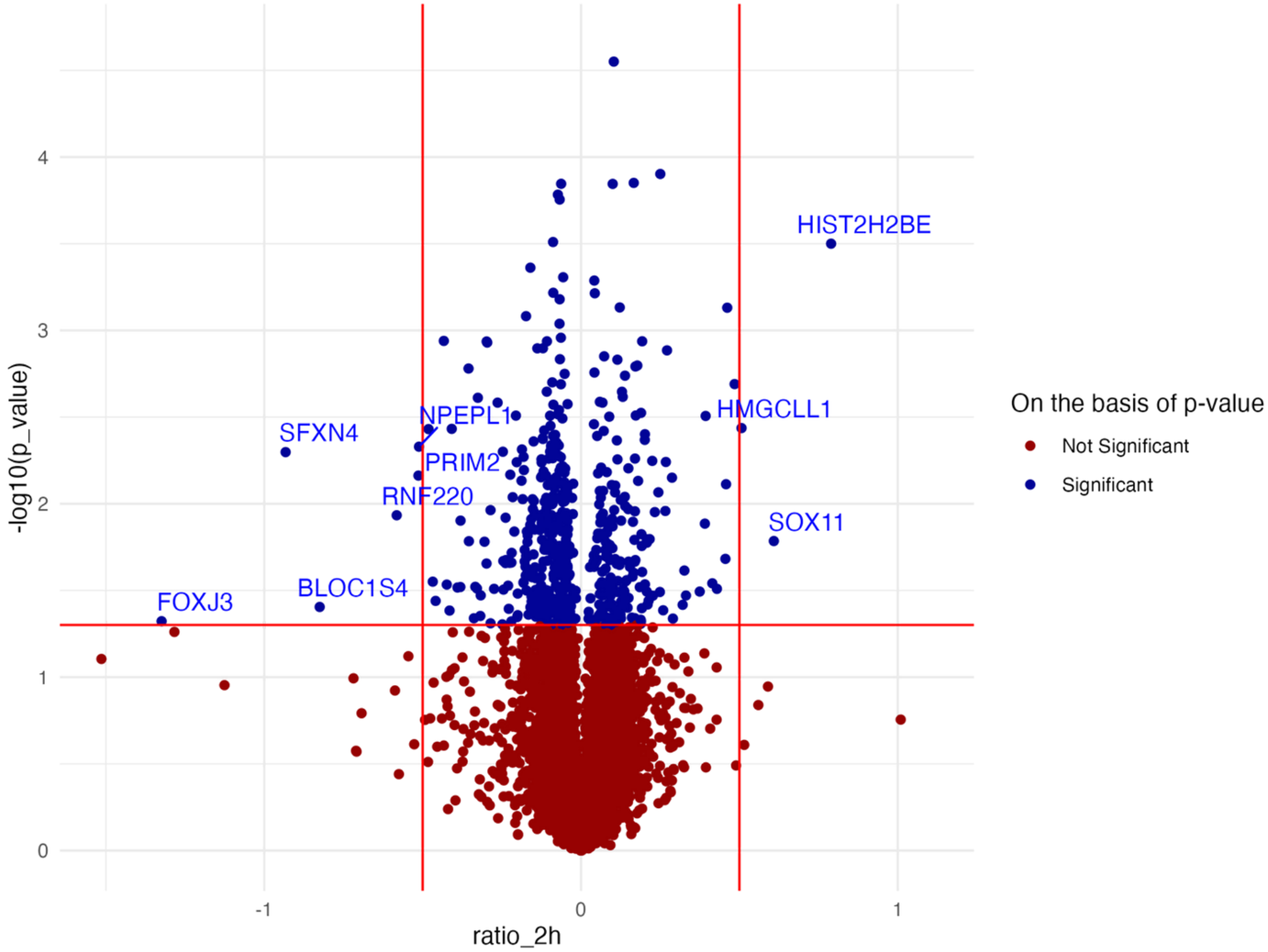
We clearly need to amp this up!

We clearly need to amp this up!

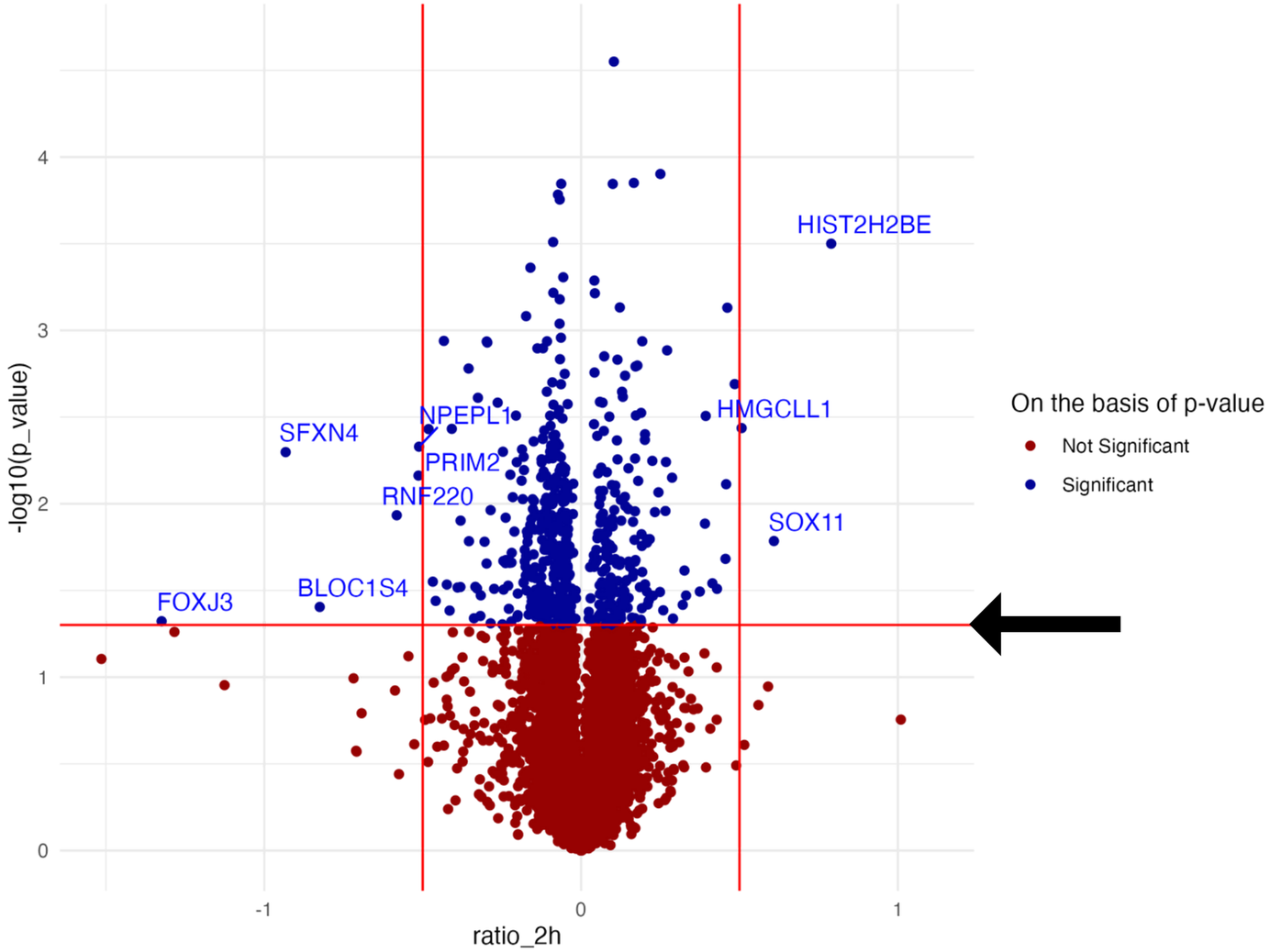
What information do we have?

- Ratio - negative says downregulation and positive ratio means upregulation!
- p-value - metric for statistical significance
- Gene symbol - allows for clear and unambiguous reference to genes

Use information to your advantage!

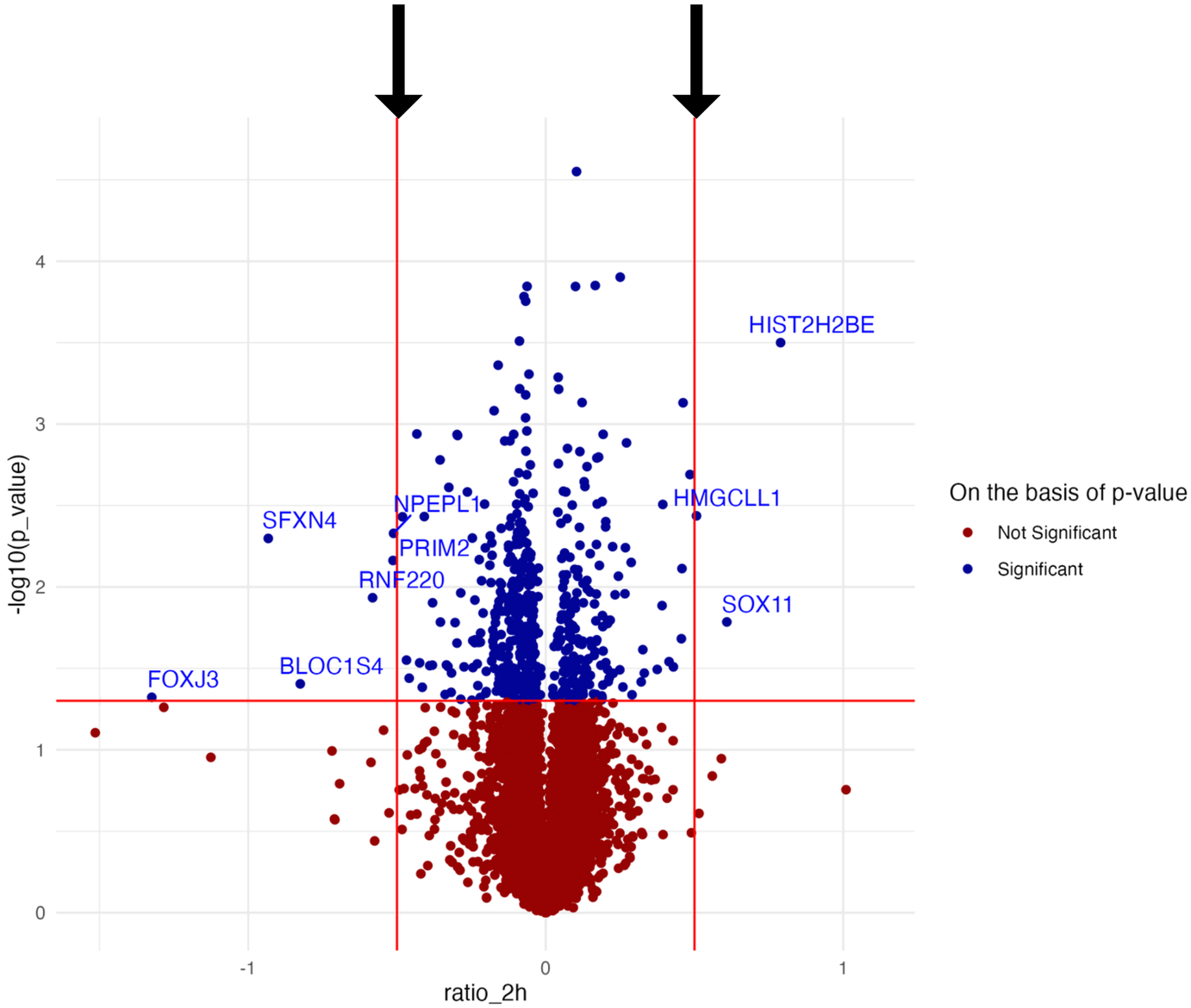


We just translated all key information in the form of a visualisation!



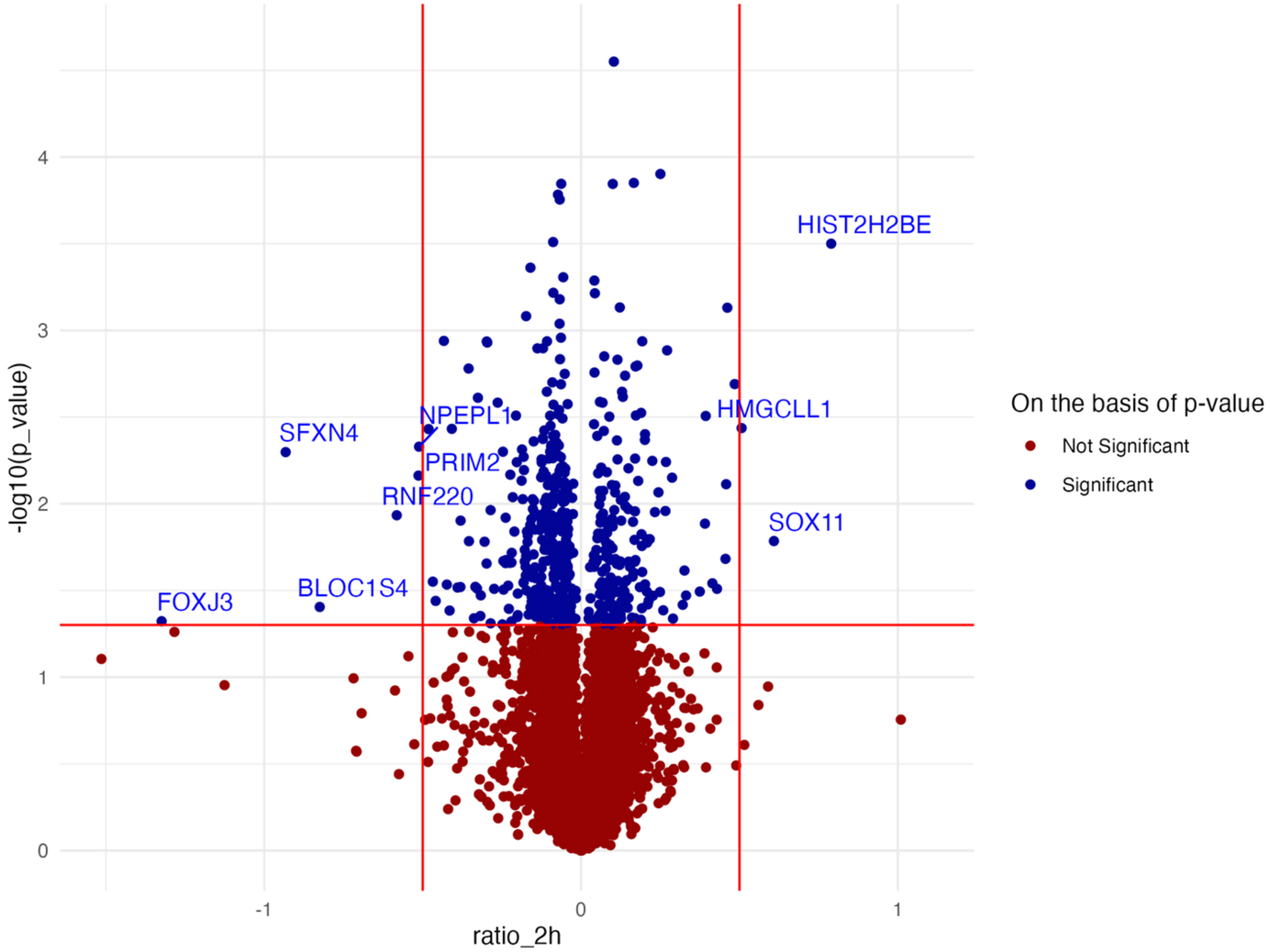
We just translated all key information in the form of a visualisation!

A threshold of 0.05 set for p-value to distinguish between significant and not significant proteins!



We just translated all key information in the form of a visualisation!

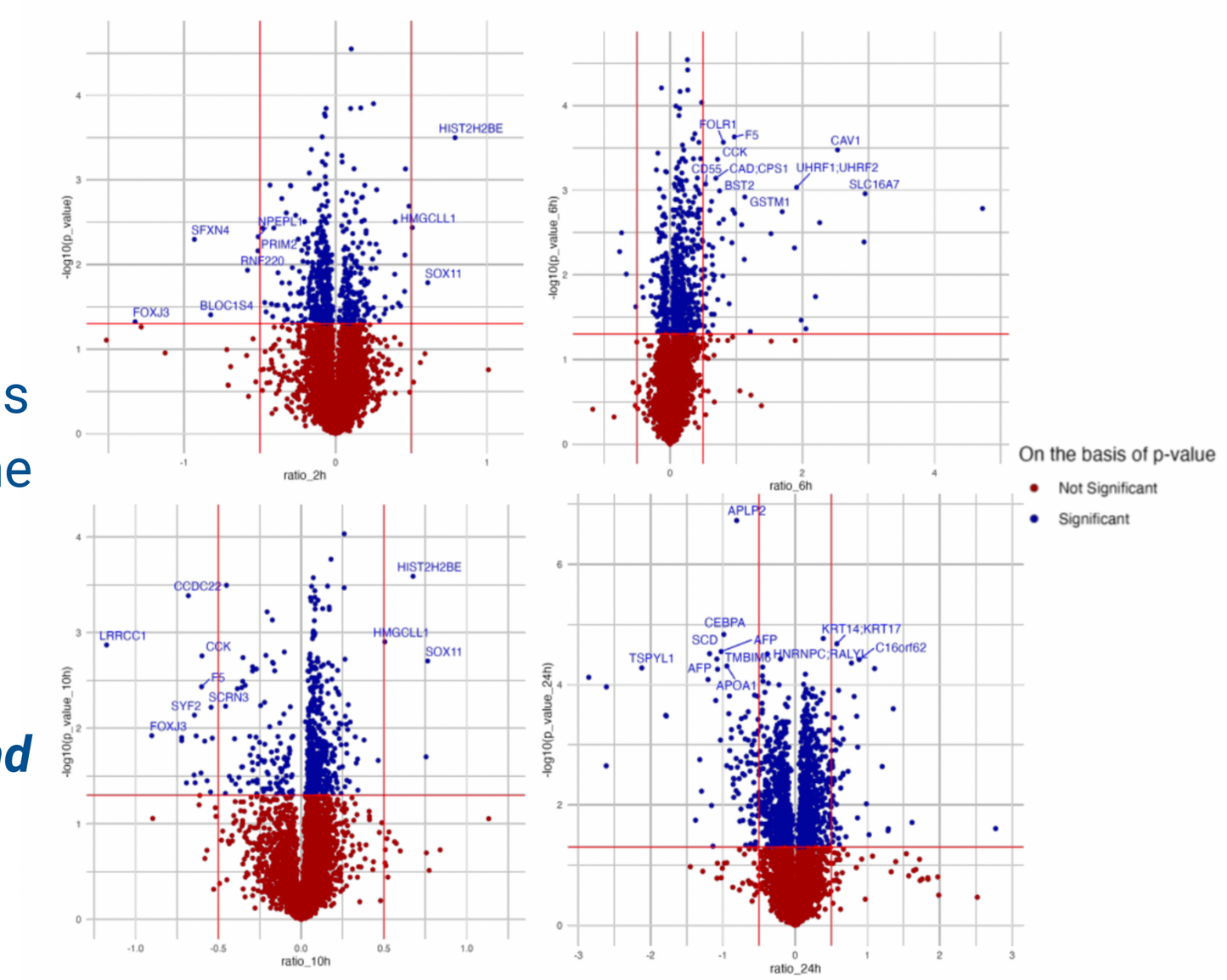
Proteins with ratio above 0.5 and below -0.5 were considered upregulated and downregulated respectively



We just translated all key information in the form of a visualisation!

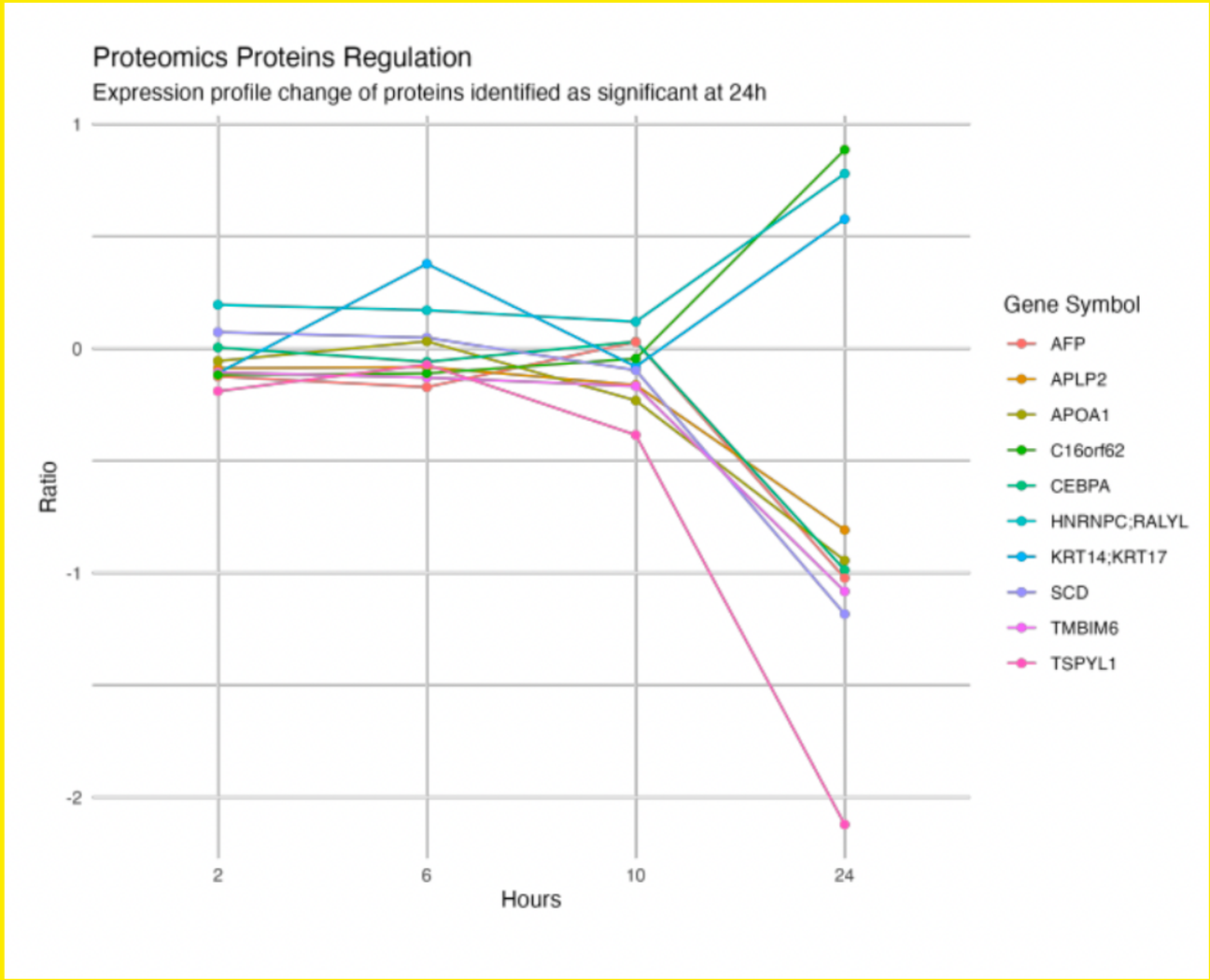
Significant proteins were sorted in ascending order based on p-value to get the top 10 proteins - which have been annotated!

- When we performed the same process for proteins at all time points, we saw that the significant proteins are not common across the four time points.
- And hence, we chose to investigate the significant proteins at 24h to ***understand later stages of infection.***

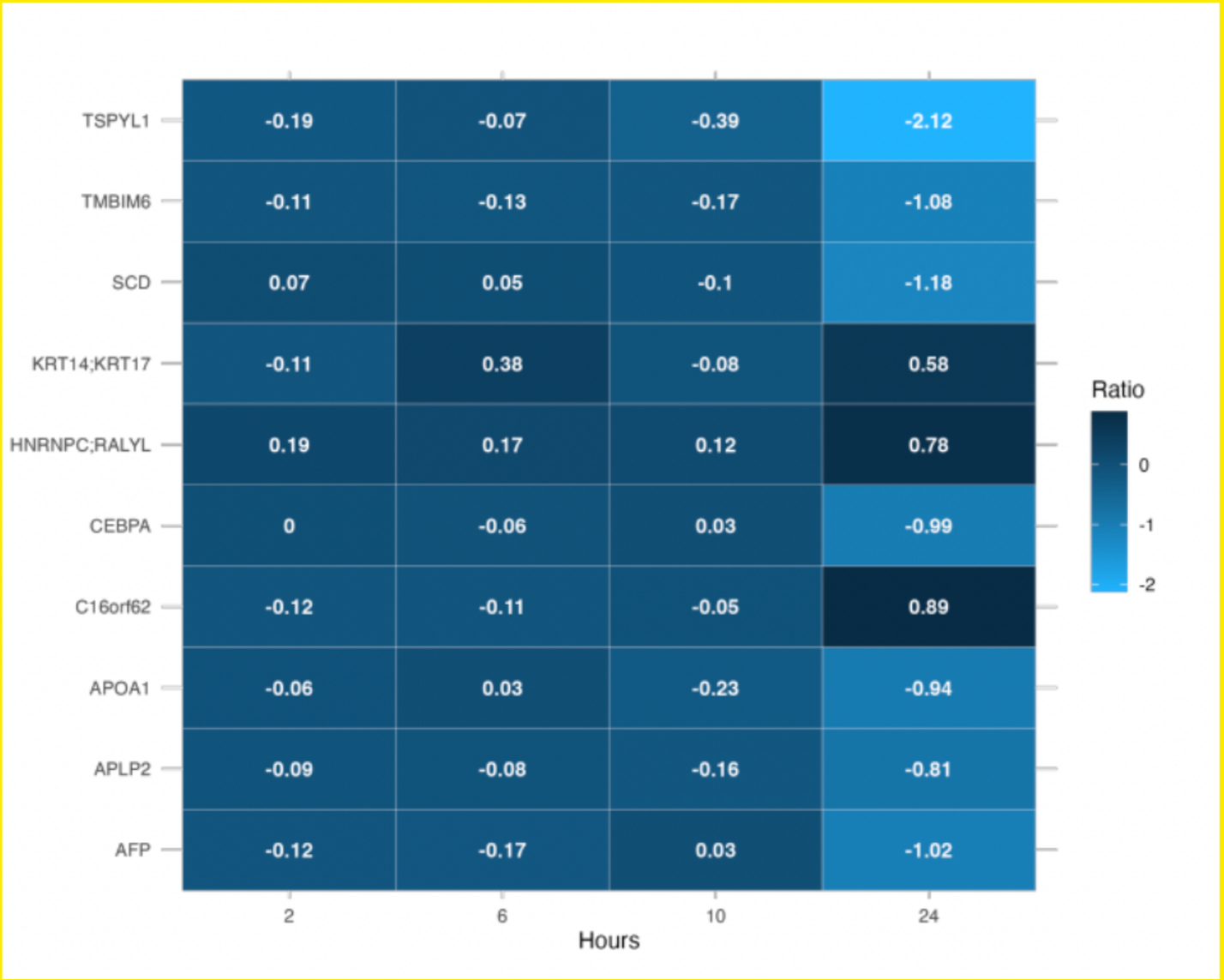


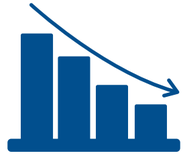
The same information can be conveyed in different formats!

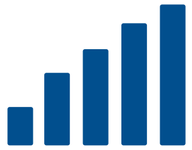
For example - studying how significant proteins at the 24h time point changed over previous hours - in the form a line plot!



Or a heatmap!


TSPYL1
TMBIM6
SCD
KRT14;KRT17
HNRNPC;RALYL
CEBPA
C16orf62
APOA1
APLP2
AFP
-0.19 -0.07 -0.39 -2.12
-0.11 -0.13 -0.17 -1.08
0.07 0.05 -0.1 -1.18
-0.11 0.38 -0.08 0.58
0.19 0.17 0.12 0.78
0 -0.06 0.03 -0.99
-0.12 -0.11 -0.05 0.89
-0.06 0.03 -0.23 -0.94
-0.09 -0.08 -0.16 -0.81
-0.12 -0.17 0.03 -1.02
2
6
10
24
Hours
Ratio
0
-1
-2





Conclusions

- **Purpose is key** - you need to have a clear idea of what you need to convey/represent
- **Start small** and then increase complexity - and always have a backup plan!
- Focus on mastering one **visualisation tool** and then dabble with more
- **Experiment** with colours, annotations, chart types.
- **Practice** making complex and interactive visualtions!

Resources

- **TidyTuesday Project:** TidyTuesday is a weekly social data project. All are welcome to participate!
- **From Data to Viz:** an excellent repository of chart types based on input data format which comes in the form of a decision tree leading to a set of potentially appropriate visualizations to represent the dataset.
- **A ggplot2 Tutorial for Beautiful Plotting in R by Cedric Scherer:** An extensive tutorial containing a general introduction to **ggplot2** as well as many examples of how to modify a **ggplot**, step by step.
- **#30DayChartChallenge:** A community-driven event with the goal to create a data visualization on a certain topic each day of April.

Mansi Aggarwal

Data Science @ Monash | Leadership Fellow @ WWCode Data Science

Thank you!