DS 203
COURSE OVERVIEW

Semester 1
2024-2025

Course Instructor

Vinay Kulkarni

vinayk@iitb.ac.in

Preferred e-mail for communication

DS 203

ds203.2023@gmail.com

DS 203

- Course title: ***Programming*** for Data Science
- Course structure (L-T-P-C): 0-2-2-6
- Tutorials : Classes, typically ‘show and tell’
- Practicals: Assignment based
- *Guest Lectures: 1 – 2 session(s) by experienced Professionals*

- Assignments : About 10 … major and minor
- Self study : 5 hours per week

DS 203

Assignments	• 8 - 10 • Every submission will be reviewed for completeness and correctness • There will be penalties for late / no / fraudulent submissions
Evaluation scheme	• 10% : 2 surprise quiz • 30% : Mid-semester test • 30% : Project (Group Project, max 4 members) • 30% : End-semester
Penalties	• (-1) : Late / non submission of assignments (each event) • (-10) : Copying assignments / project (each event) • (-10) : Fraudulent assignment submissions (each event) • The penalty will be in addition to zero credit (where applicable) for that particular submission
Attendance	• As per IITB rules for attendance (self regulated) • No attendance will be taken in class, except during quiz / test / examination

Quiz / Test / Exam

- SAFE app, Forms (Google / PDF), Moodle may be used
- **Paper** option will always be available as a standby
- Tests and exams
  - Will involve reasoning, programming
  - Will involve **application** of the knowledge and skill learnt through attending classes and doing the assignments
  - Will be, **most likely**, open – including internet access

SAFE Application

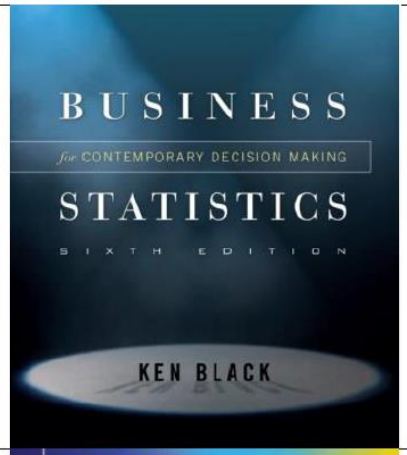
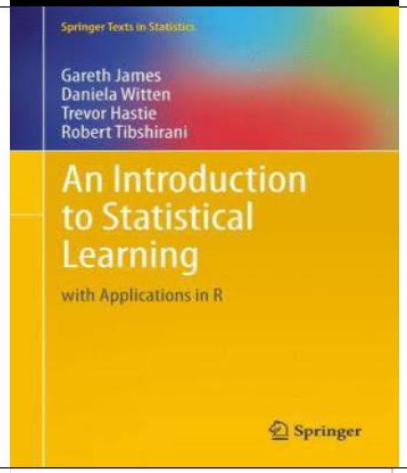
- SAFE app will be used for conducting quiz, and for marking attendance during the quiz
- All participants **should** compulsorily register themselves on SAFE app using the following registration code:

20FLXYSI

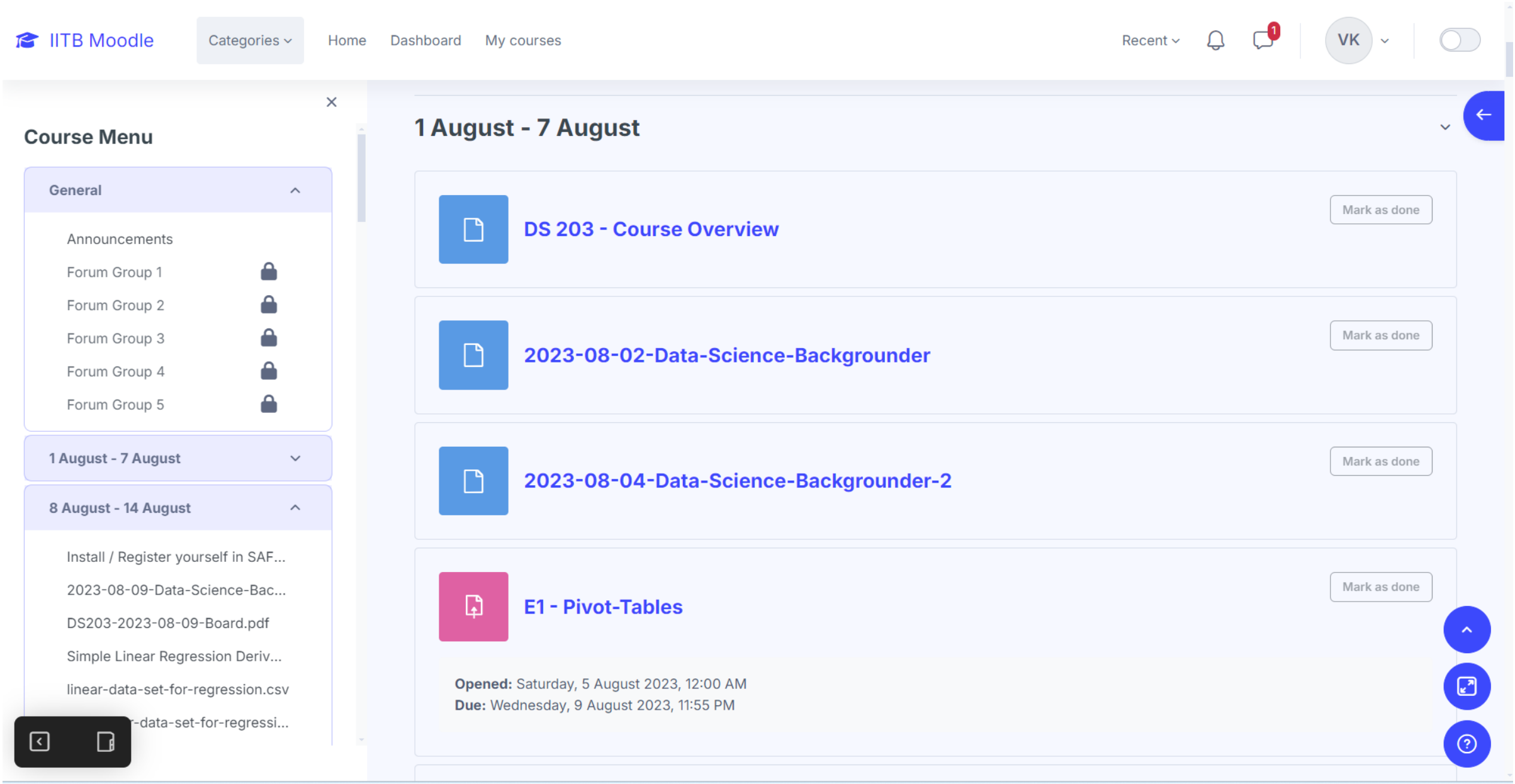
- You may mark your class attendance using the app (for your own records)
- In case of teething troubles, write to safe@iitb.ac.in and visit their office in the CC to get them resolved

Books and References

- Learning Data Science : https://learningds.org/intro.html
- Veridical Data Science : https://vdsbook.com/

For topics related to Statistics	Business Statistics for Contemporary Decision Making Author: Ken Black (Available online – 6th Edition)	
For topics related to Machine Learning	An Introduction to Statistical Learning Authors: Gareth James and others (Available online)	

Course Assets on Moodle



Data Science

What is Data Science?

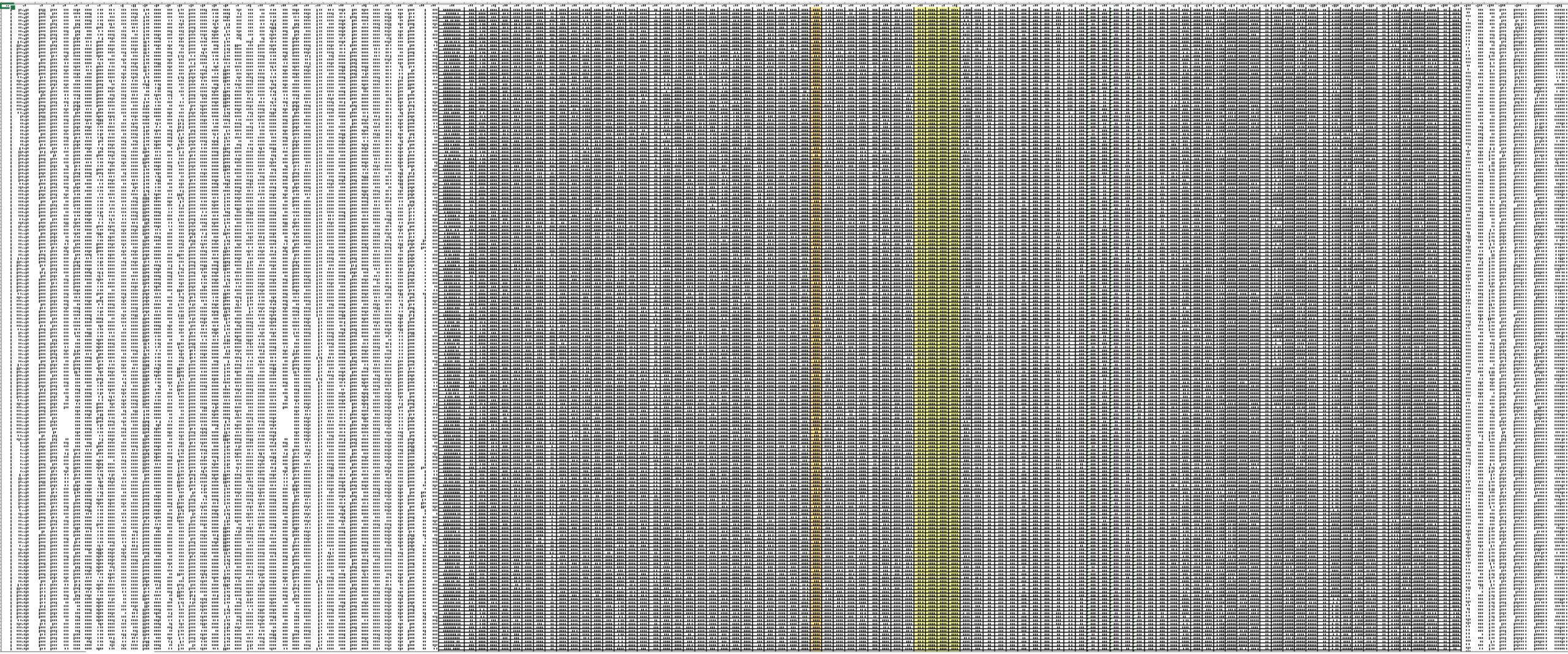
- Data science is a **broad field** that encompasses the overall process of **extracting insights** and **knowledge** from data. It involves **collecting, cleaning, organizing, analyzing, and interpreting data** to uncover **patterns, trends, and meaningful information**.
- Data science utilizes various techniques, methodologies, and tools to **extract valuable insights** from data.

Case Study

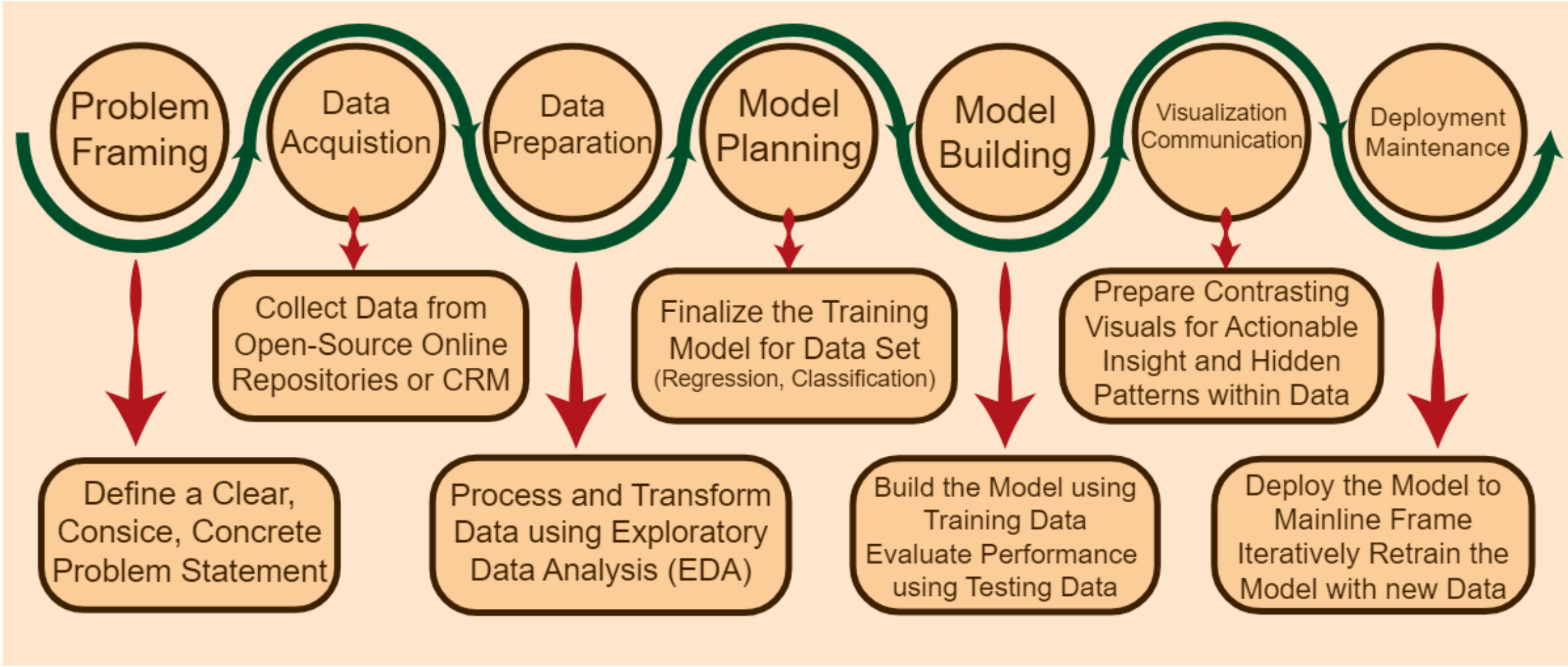
Case Study Overview

Attribute	Details
The problem	Optimization of process(es) in a chemical plant
Goals	1. Operating parameters' optimization 2. Throughput prediction 3. Breakdown prediction 4. Predictive Maintenance of equipment
Basis	Multiple year, daily averages of operational parameters available from Data Acquisition System
Columns	132
Rows	2657 (approximately 7 years' daily averages data)

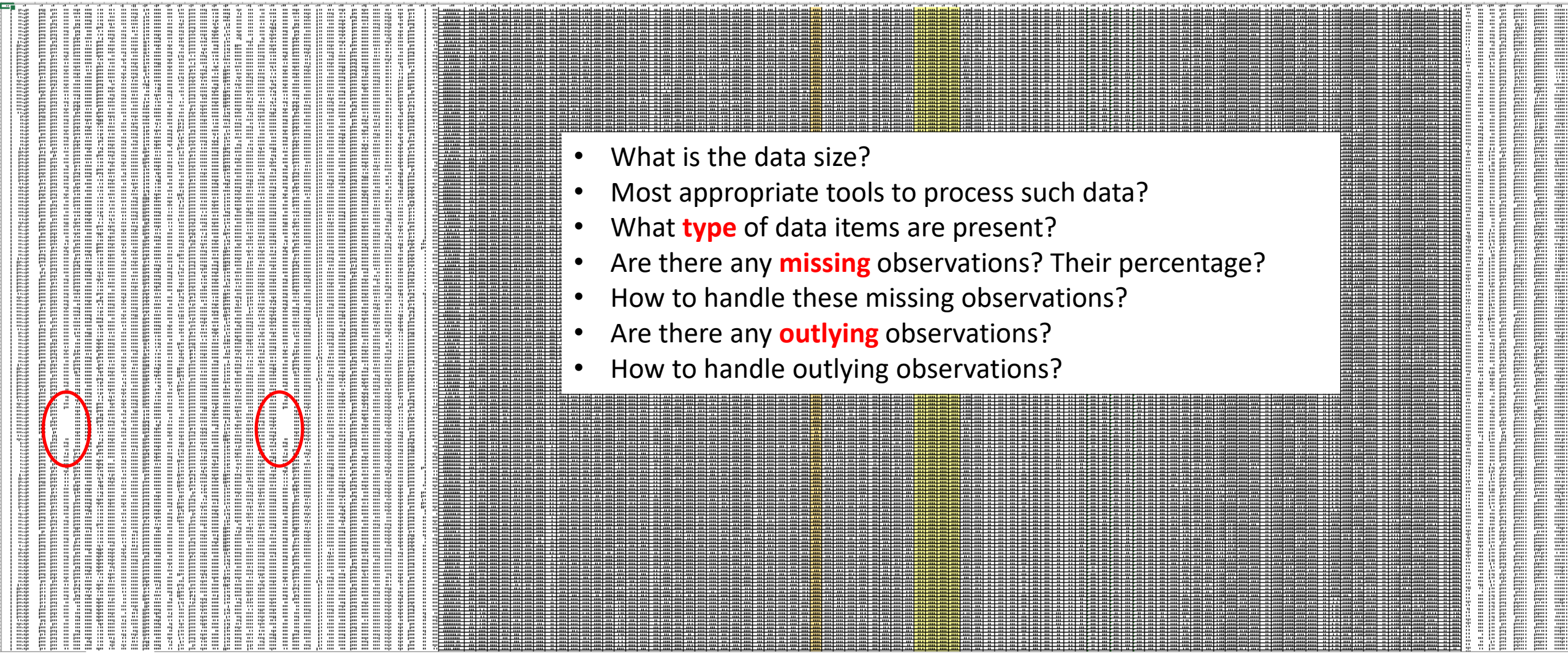
The Data …



The Data Science Process

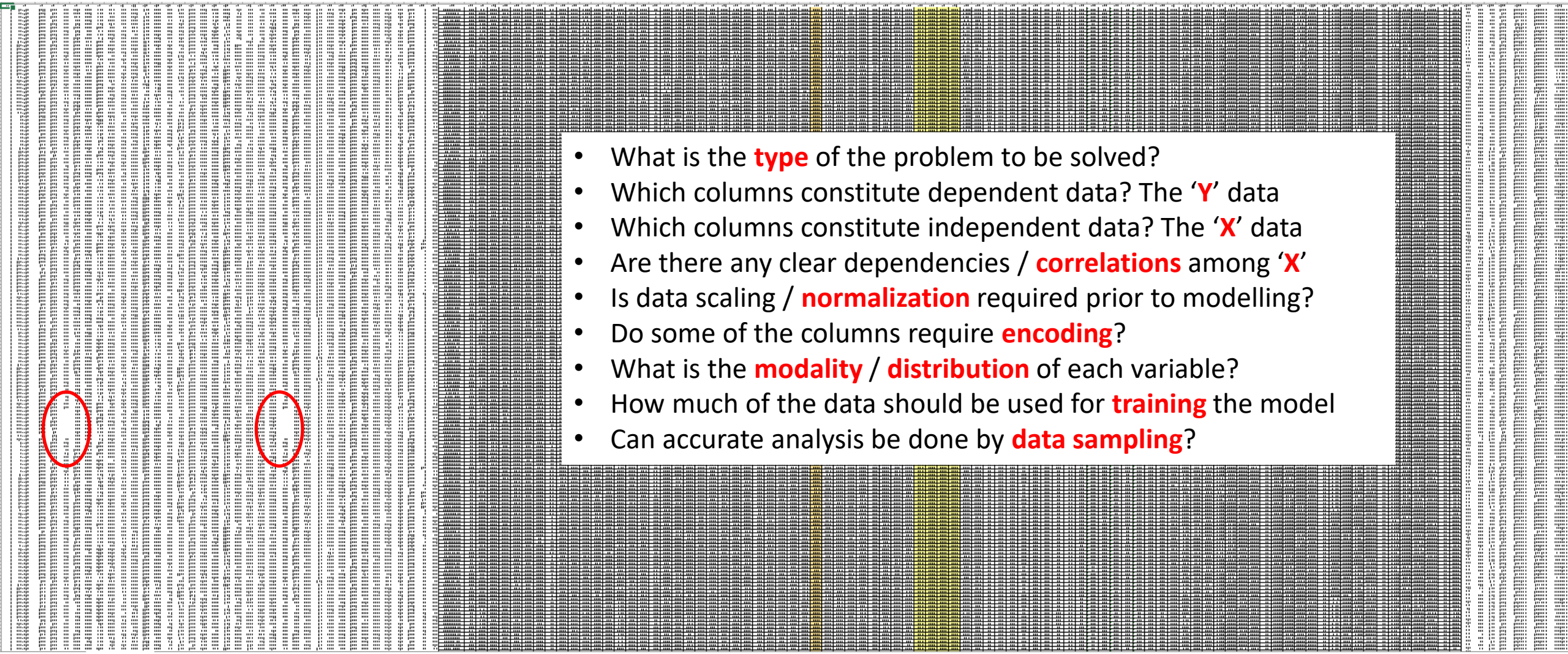


Preliminary analysis of the Data



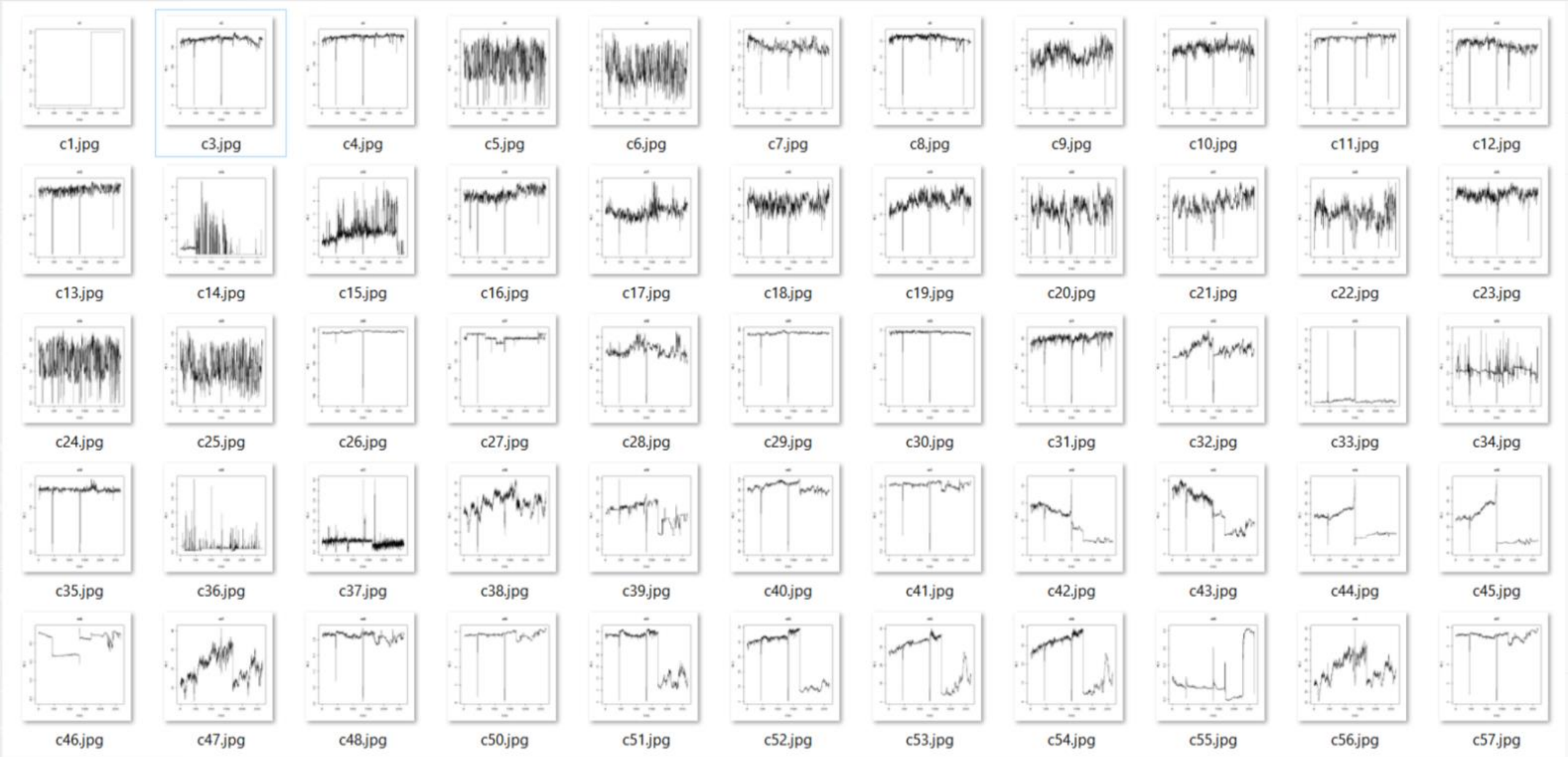
- What is the data size?
- Most appropriate tools to process such data?
- What **type** of data items are present?
- Are there any **missing** observations? Their percentage?
- How to handle these missing observations?
- Are there any **outlying** observations?
- How to handle outlying observations?

Early analysis / decisions based on Data

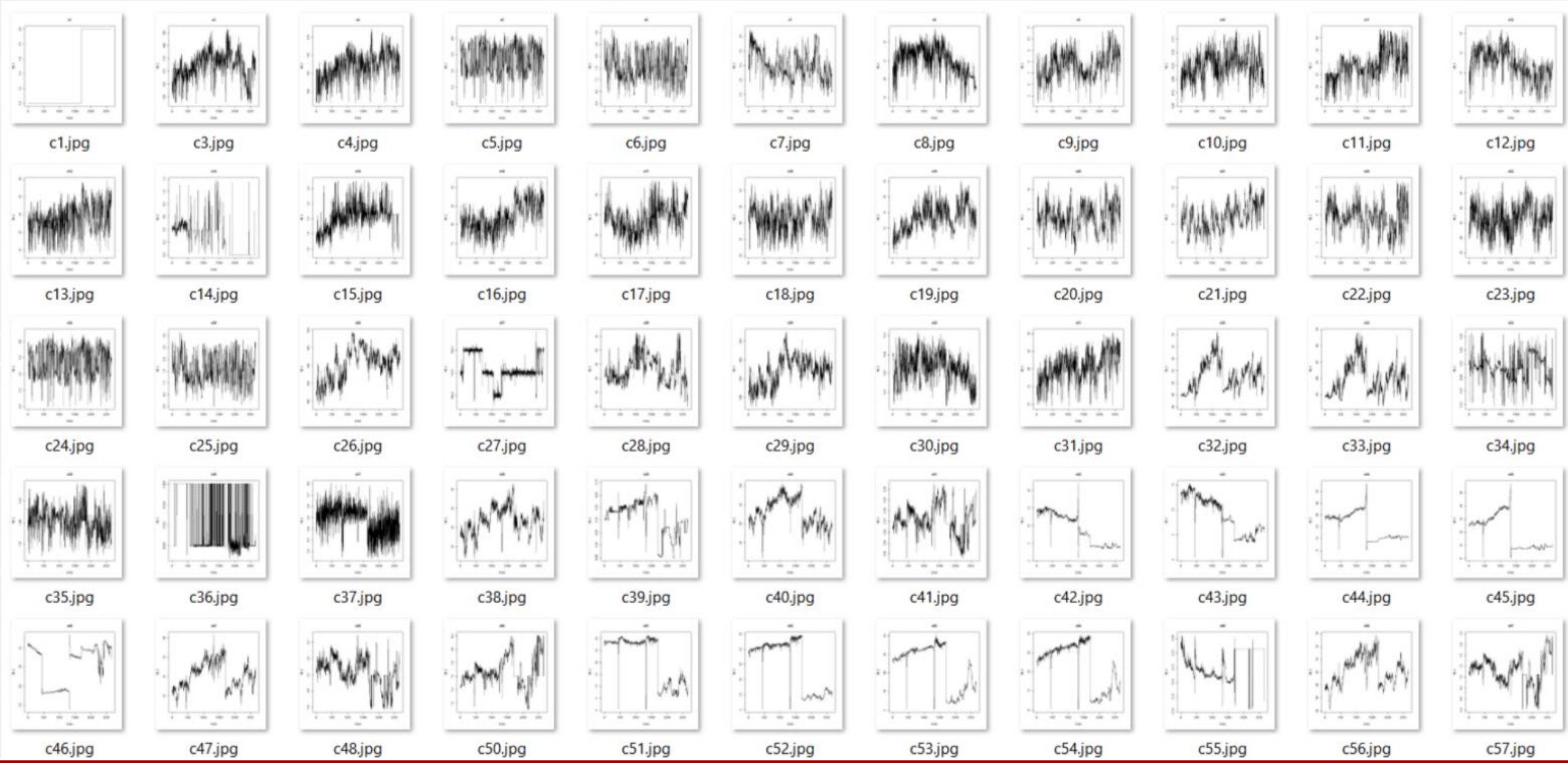


- What is the **type** of the problem to be solved?
- Which columns constitute dependent data? The '**Y**' data
- Which columns constitute independent data? The '**X**' data
- Are there any clear dependencies / **correlations** among '**X**'
- Is data scaling / **normalization** required prior to modelling?
- Do some of the columns require **encoding**?
- What is the **modality** / **distribution** of each variable?
- How much of the data should be used for **training** the model
- Can accurate analysis be done by **data sampling**?

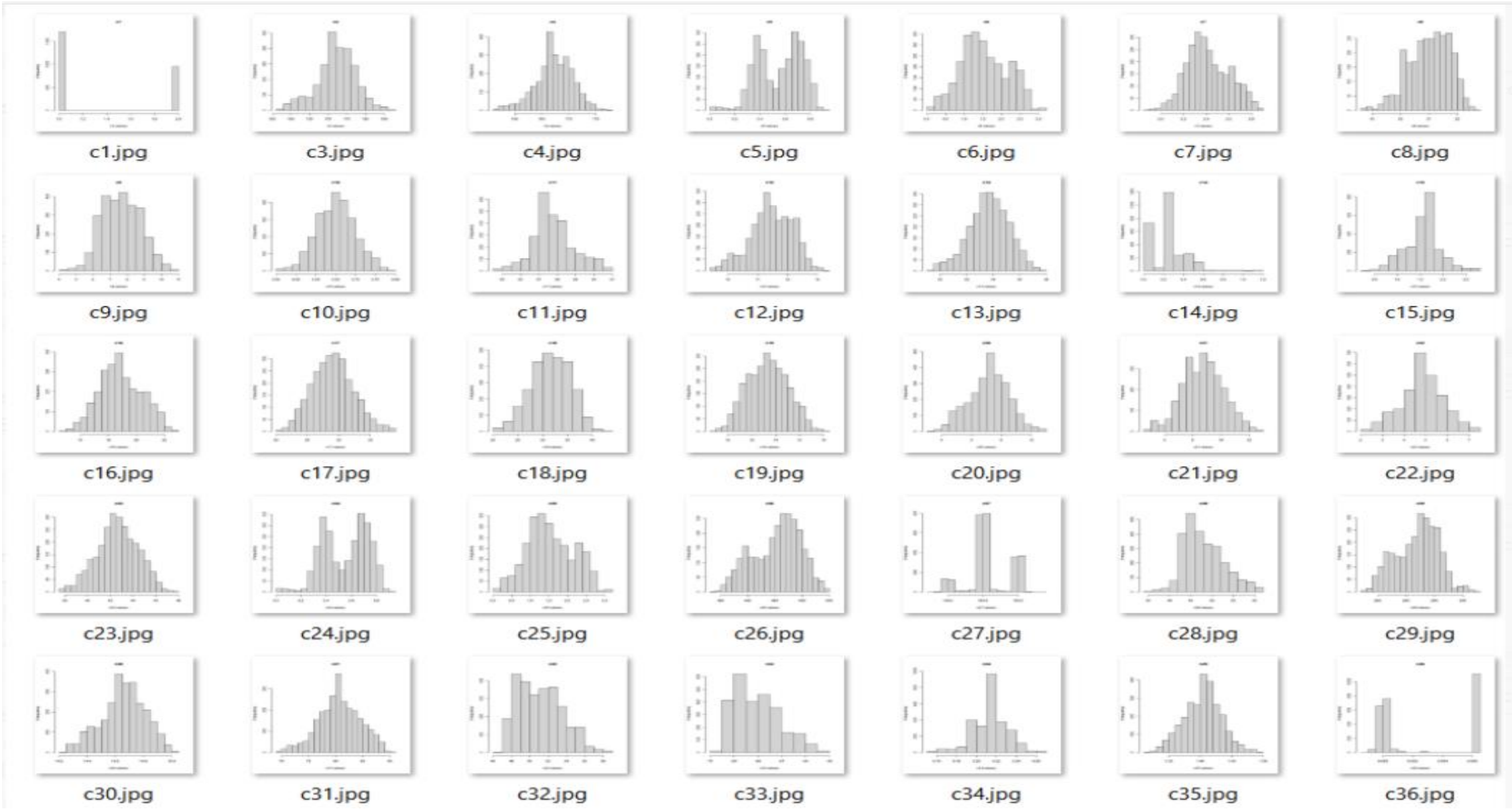
Data plots to understand trends / detect outliers



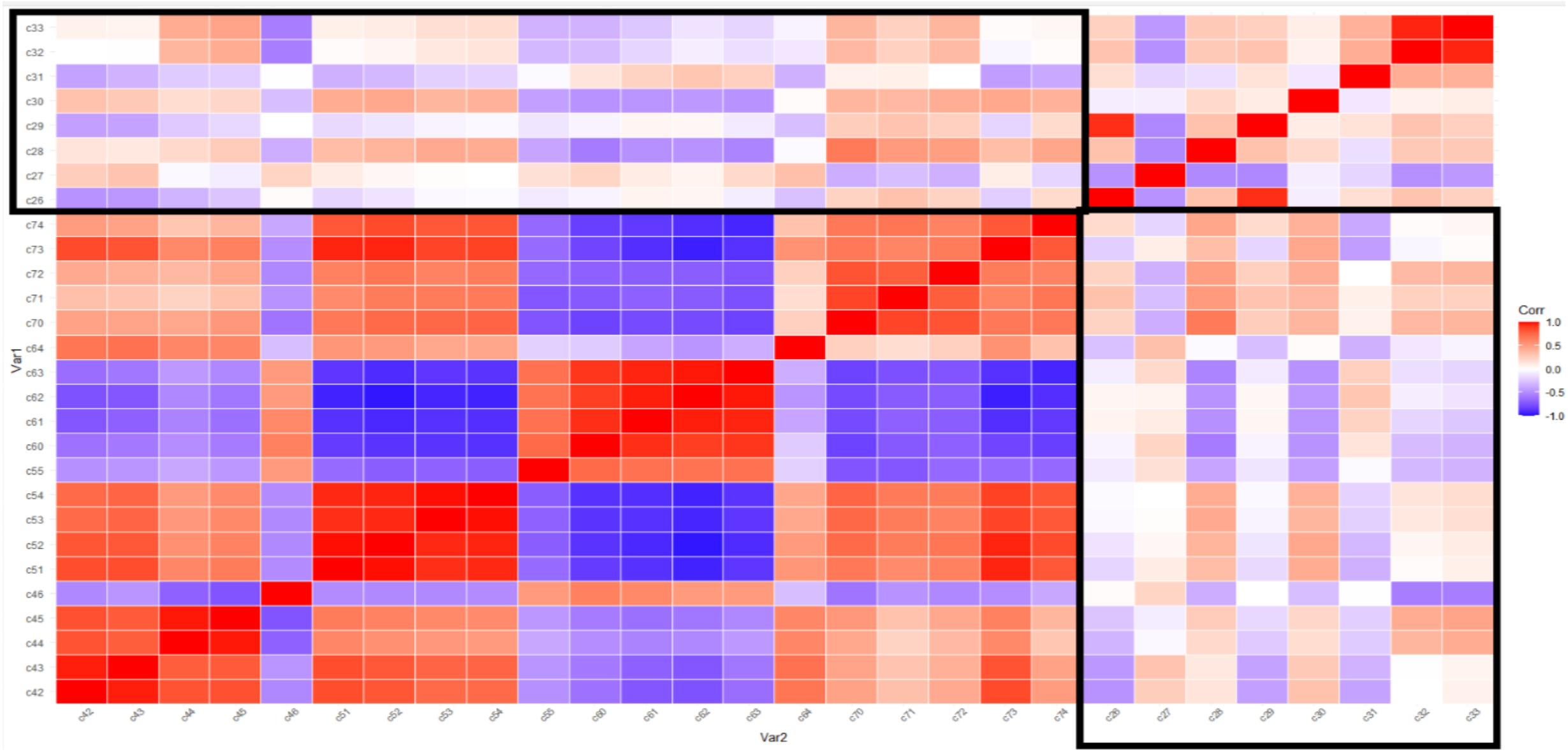
Data plots after **processing** outliers



Modality and Distributions



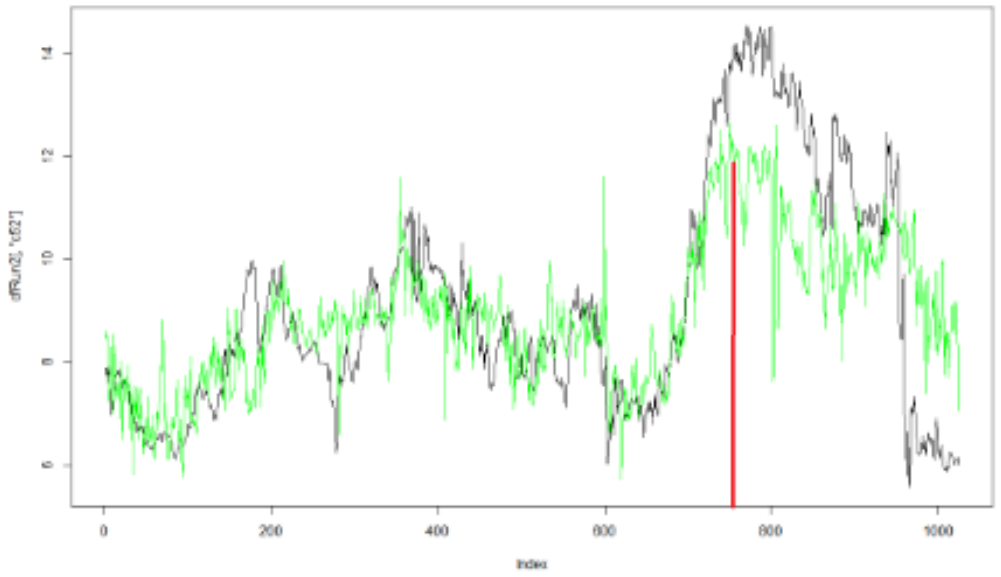
Correlation between variables



Use of Linear Regression to **model** and **predict** a parameter

Training: First 750 data points

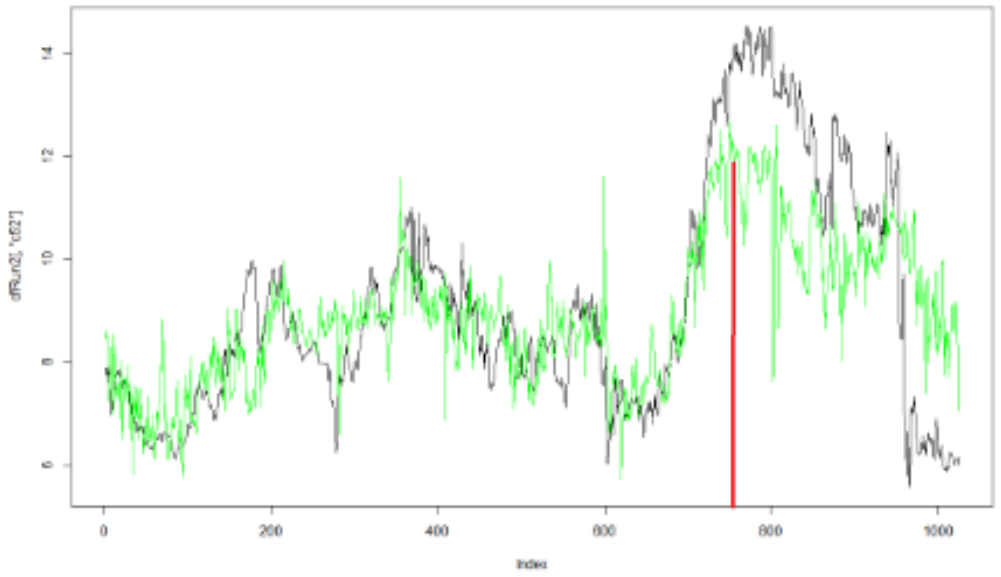
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -16.209642   5.592980  -2.898 0.003866 **
c161          0.011909   0.001154  10.317  < 2e-16 ***
c28           0.192089   0.031109   6.175 1.10e-09 ***
c7            1.762212   0.244322   7.213 1.38e-12 ***
c17          -0.099812   0.019082  -5.231 2.21e-07 ***
c158          0.123261   0.018093   6.813 2.01e-11 ***
c160          0.004787   0.001771   2.703 0.007041 **
c39           6.898476   1.100166   6.270 6.18e-10 ***
c22          -0.104352   0.034339  -3.039 0.002460 **
c11          -0.122940   0.036342  -3.383 0.000756 ***
c15          -0.371219   0.056582  -6.561 1.02e-10 ***
c30           1.702216   0.337681   5.041 5.85e-07 ***
c23          -0.274084   0.039142  -7.002 5.74e-12 ***
c35           5.033540   1.451099   3.469 0.000554 ***
c16          -0.381731   0.073873  -5.167 3.07e-07 ***
c139         -0.221037   0.036769  -6.011 2.91e-09 ***
c31           0.156248   0.022446   6.961 7.55e-12 ***
c143         -0.236297   0.034607  -6.828 1.82e-11 ***
c157          0.160972   0.038140   4.221 2.75e-05 ***
c163          0.009764   0.002779   3.513 0.000471 ***
c9           -0.267089   0.070883  -3.768 0.000178 ***
c8           -0.369612   0.122558  -3.016 0.002652 **
c10           3.208038   1.541097   2.082 0.037723 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8832 on 727 degrees of freedom
Multiple R-squared:  0.6309,    Adjusted R-squared:  0.6198
F-statistic: 56.49 on 22 and 727 DF,  p-value: < 2.2e-16
```



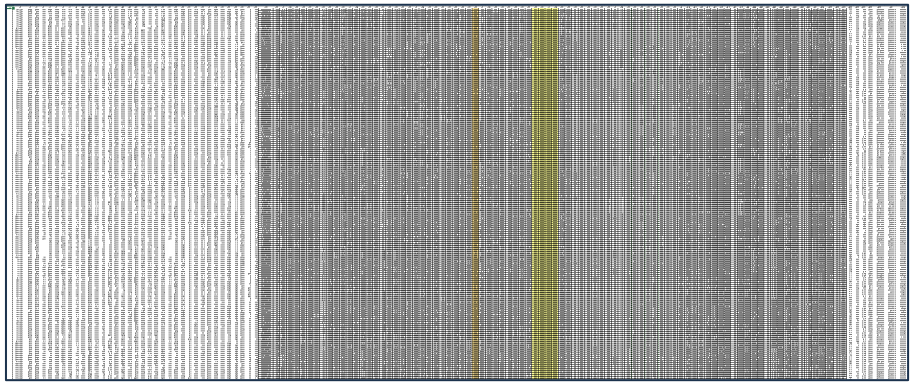
Interpretation of results

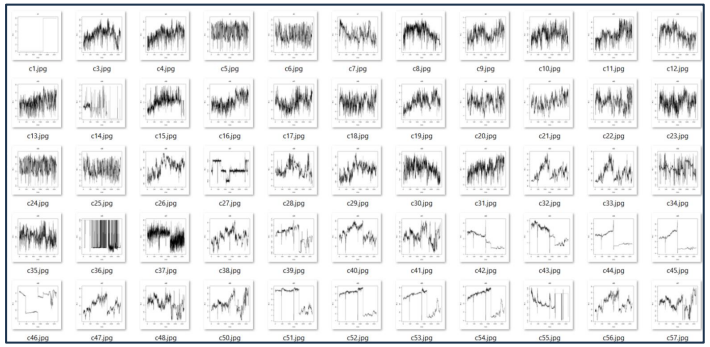
Training: First 750 data points

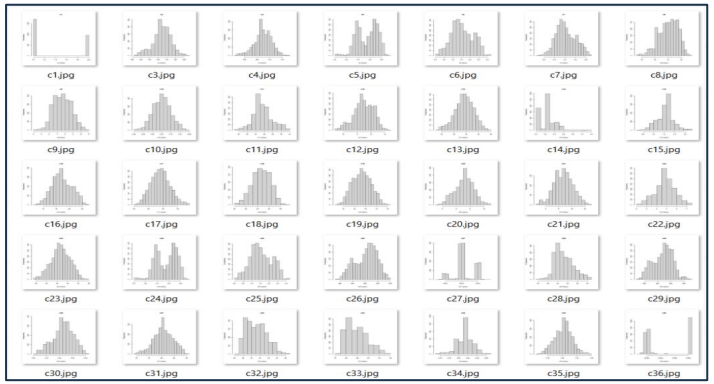
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -16.209642   5.592980  -2.898 0.003866 **
c161          0.011909   0.001154  10.317  < 2e-16 ***
c28           0.192089   0.031109   6.175 1.10e-09 ***
c7            1.762212   0.244322   7.213 1.38e-12 ***
c17          -0.099812   0.019082  -5.231 2.21e-07 ***
c158          0.123261   0.018093   6.813 2.01e-11 ***
c160          0.004787   0.001771   2.703 0.007041 **
c39           6.898476   1.100166   6.270 6.18e-10 ***
c22          -0.104352   0.034339  -3.039 0.002460 **
c11          -0.122940   0.036342  -3.383 0.000756 ***
c15          -0.371219   0.056582  -6.561 1.02e-10 ***
c30           1.702216   0.337681   5.041 5.85e-07 ***
c23          -0.274084   0.039142  -7.002 5.74e-12 ***
c35           5.033540   1.451099   3.469 0.000554 ***
c16          -0.381731   0.073873  -5.167 3.07e-07 ***
c139         -0.221037   0.036769  -6.011 2.91e-09 ***
c31           0.156248   0.022446   6.961 7.55e-12 ***
c143         -0.236297   0.034607  -6.828 1.82e-11 ***
c157          0.160972   0.038140   4.221 2.75e-05 ***
c163          0.009764   0.002779   3.513 0.000471 ***
c9           -0.267089   0.070883  -3.768 0.000178 ***
c8           -0.369612   0.122558  -3.016 0.002652 **
c10           3.208038   1.541097   2.082 0.037723 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8832 on 727 degrees of freedom
Multiple R-squared:  0.6309,    Adjusted R-squared:  0.6198
F-statistic: 56.49 on 22 and 727 DF,  p-value: < 2.2e-16
```

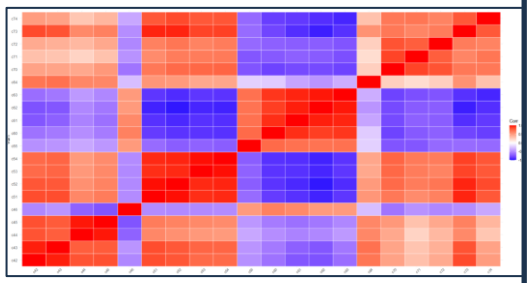


Goal of Data Science : **Extract Insights from Data**









Training: First 750 data points

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -16.209642   5.592980  -2.898 0.003866 **
c161          0.011909   0.001154  10.317  < 2e-16 ***
c28           0.192089   0.031109   6.175 1.10e-09 ***
c7            1.762212   0.244322   7.213 1.38e-12 ***
c17          -0.099812   0.019082  -5.231 2.21e-07 ***
c158          0.123261   0.018093   6.813 2.01e-11 ***
c160          0.004787   0.001771   2.703 0.007041 **
c39           6.898476   1.100166   6.270 6.18e-10 ***
c22          -0.104352   0.034339  -3.039 0.002460 **
c11          -0.122940   0.036342  -3.383 0.000756 ***
c15          -0.371219   0.056582  -6.561 1.02e-10 ***
c30           1.702216   0.337681   5.041 5.85e-07 ***
c23          -0.274084   0.039142  -7.002 5.74e-12 ***
c35           5.033540   1.451099   3.469 0.000554 ***
c16          -0.381731   0.073873  -5.167 3.07e-07 ***
c139         -0.221037   0.036769  -6.011 2.91e-09 ***
c31           0.156248   0.022446   6.961 7.55e-12 ***
c143         -0.236297   0.034607  -6.828 1.82e-11 ***
c157          0.160972   0.038140   4.221 2.75e-05 ***
c163          0.009764   0.002779   3.513 0.000471 ***
c9           -0.267089   0.070883  -3.768 0.000178 ***
c8           -0.369612   0.122558  -3.016 0.002652 **
c10           3.208038   1.541097   2.082 0.037723 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8832 on 727 degrees of freedom
Multiple R-squared:  0.6309,    Adjusted R-squared:  0.6198
F-statistic: 56.49 on 22 and 727 DF,  p-value: < 2.2e-16
```