Assignment II

By

Parsa Besharat

An assignment’s handout submitted as part of the requirements
for the lecture, Geomodeling, of MSc Mathematics of Data and Resources Sciences
at the Technische Universität Bergakademie Freiberg

November, 2024

Supervisor: Prof. Jörg Benndorf

1. Is something unusual? Why?

The *x* and *y* columns both contain a limited range of unique values, which might indicate that these variables represent a structured grid (e.g., coordinates). This suggests that the values are not continuous but potentially rounded or categorized, which is unusual for continuous variables in natural datasets.

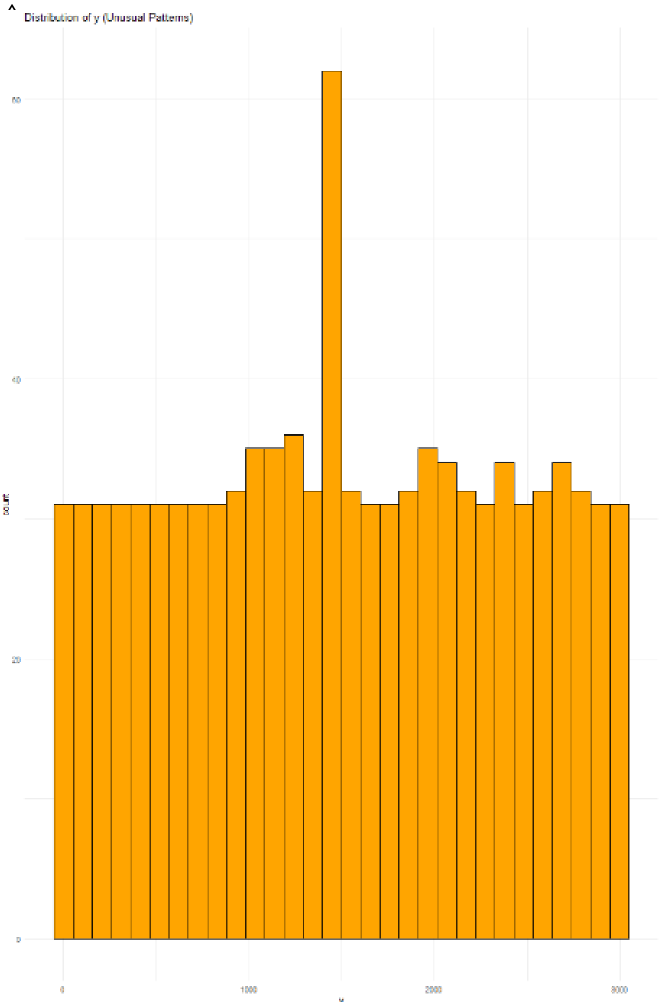

Figure 1.1, Distribution of *x*(Unusual Patterns)

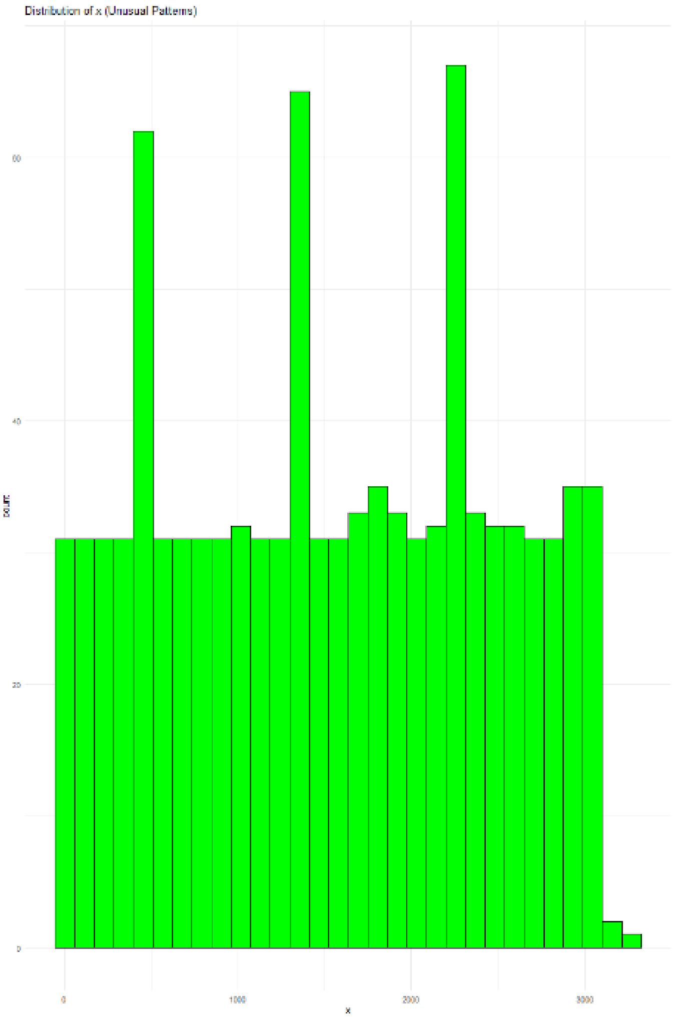

Figure 1.2, Distribution of *y* (Unusual Patterns)

2. How are the data distributed? Are the data skewed?

The distributions of *Co* and *Ni* both show right skewness, with a long tail on the higher values. The skewness values for these columns confirm that the data are not symmetrically distributed, and most data points are clustered at lower values with a few large values pulling the mean upward. This suggests that these variables may require transformation for certain types of analysis.

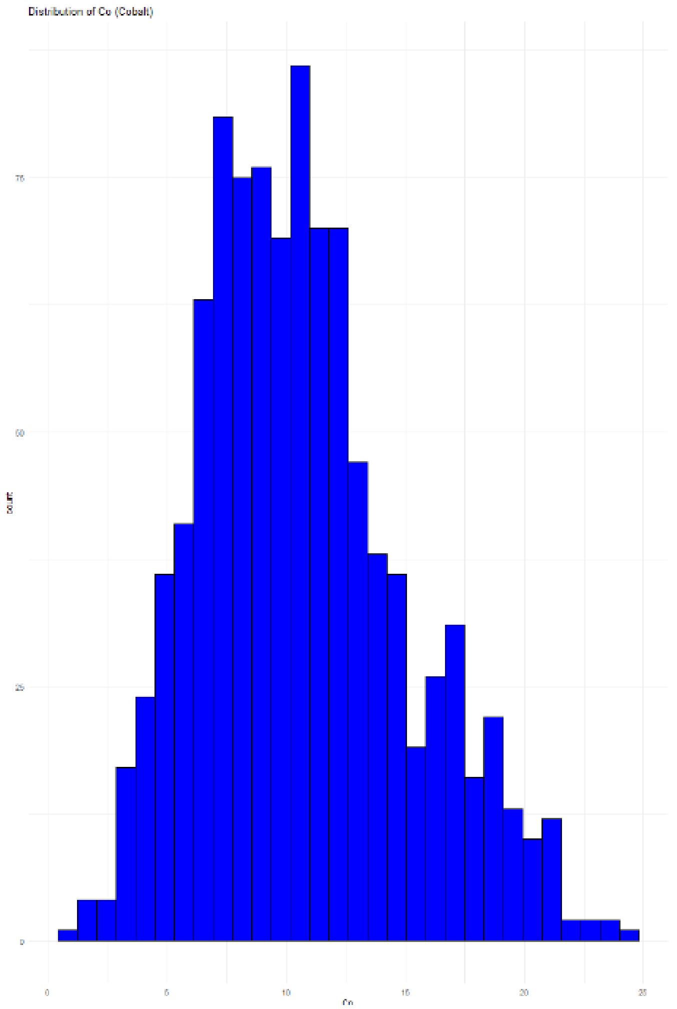

Figure 2.1, Distribution of *Co*

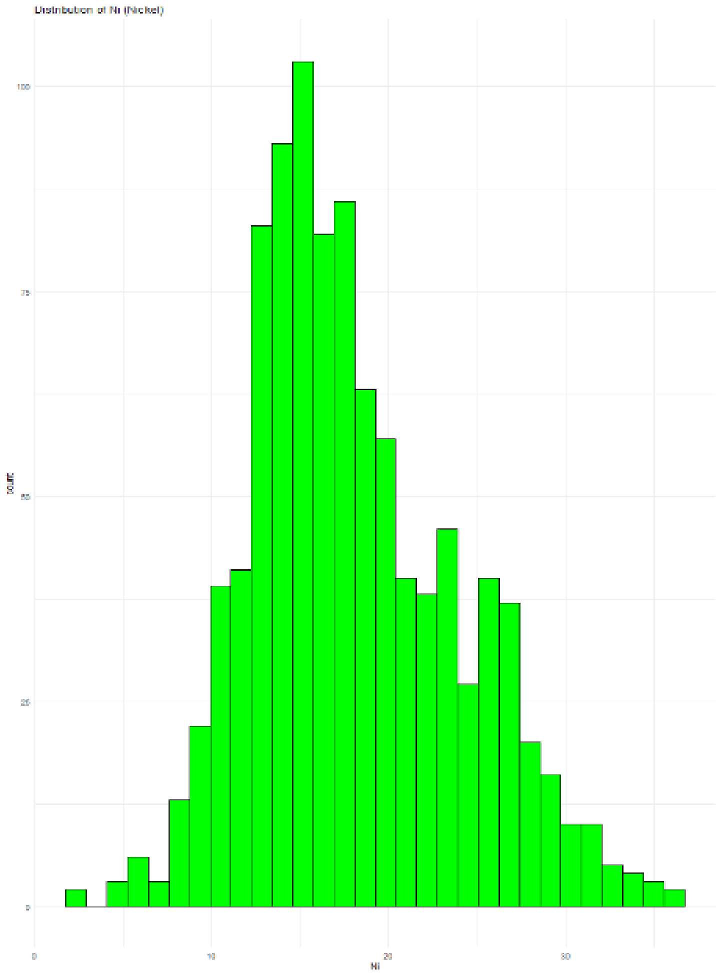

Figure 2.2, Distribution of *Ni*

3. Are there outliers or bindings (are the values rounded)?

Boxplots for *Co* and *Ni* indicate the presence of outliers, especially for larger values, which could represent exceptional cases in the data. The x and y columns have fewer unique values, suggesting that these columns might be rounded or discretized to fit a grid structure, as expected in spatial data or mapped coordinates.

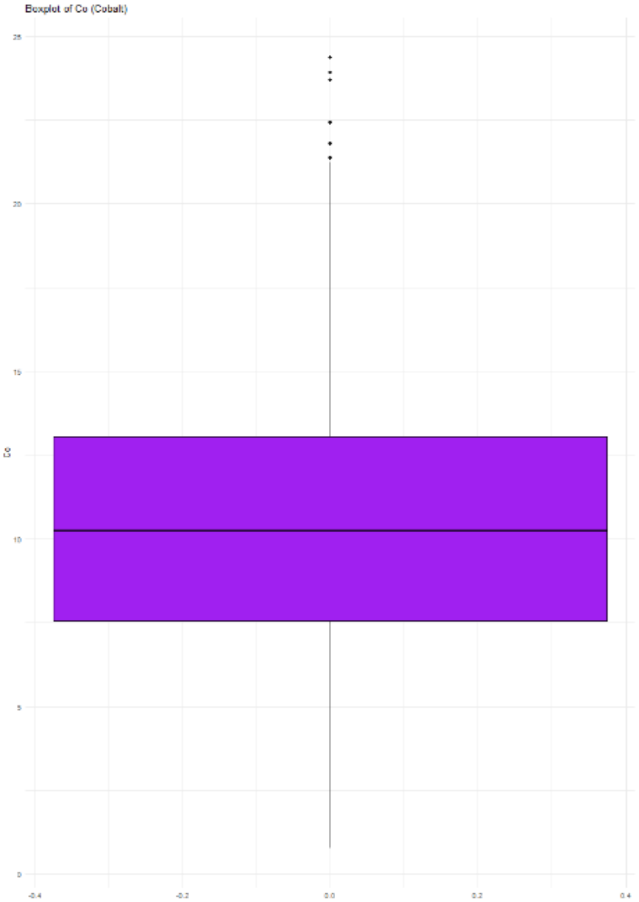

Figure 3.1, Boxplot of *Co*

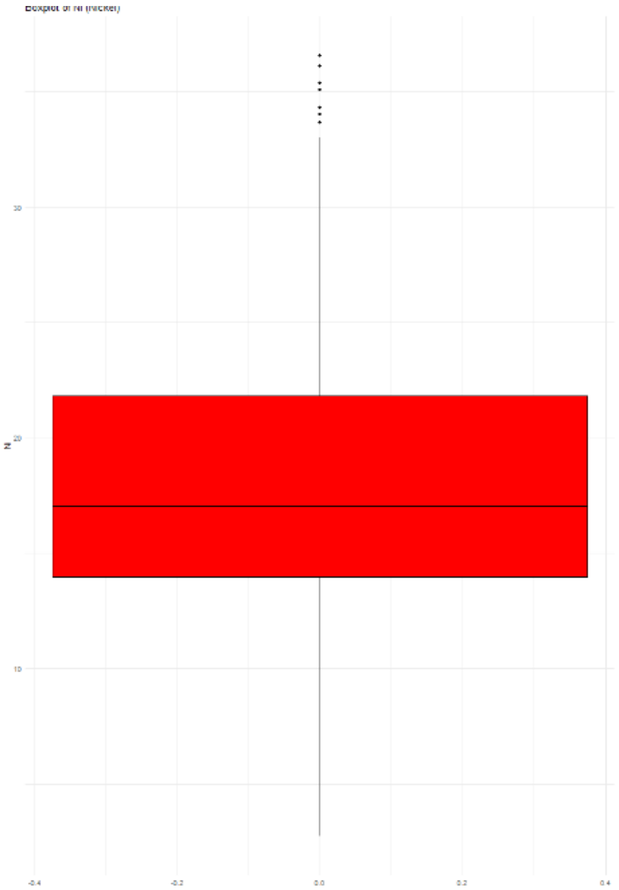

Figure 3.2, Boxplot of *Ni*

4. Is the optical impression falsified through special features?

The distribution plots for *Co* and *Ni* can sometimes give the wrong impression if the bin size is too large or small. By adjusting the number of bins in the histograms, we can more accurately represent the underlying data. In the original plots, the wide variation in the data was compressed due to coarse binning, leading to a misleading optical impression of the distribution.

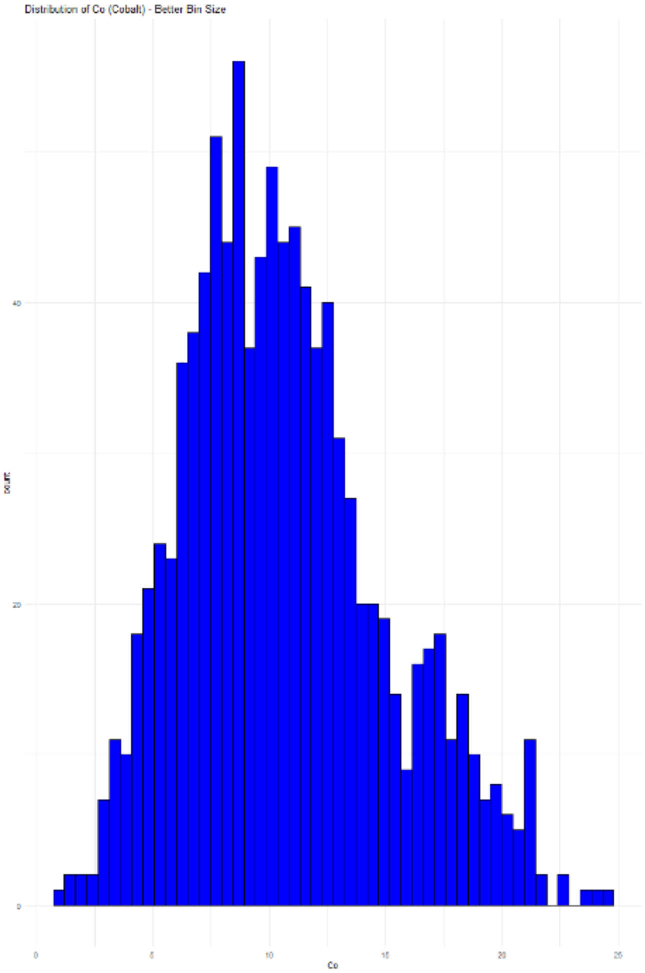

Figure 4.1, Distribution of *Co* - Better Bin Size

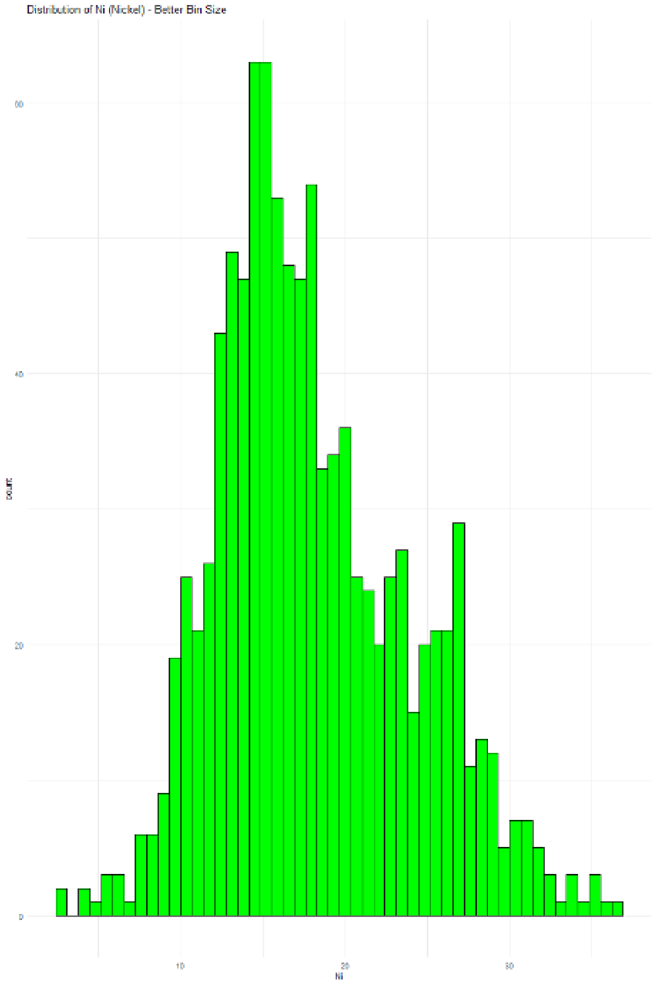

Figure 4.2, Distribution of *Ni* - Better Bin Size

5. **Can you see tectonic faults or hops in the map/data?**

The *x* and *y* columns likely represent a spatial grid, which does not show any obvious faults or disruptions. The points are evenly distributed, indicating a structured survey grid. If tectonic faults were present, we'd expect non-uniform spacing or abrupt changes in the spatial patterns, but no such features are visible in the current data.

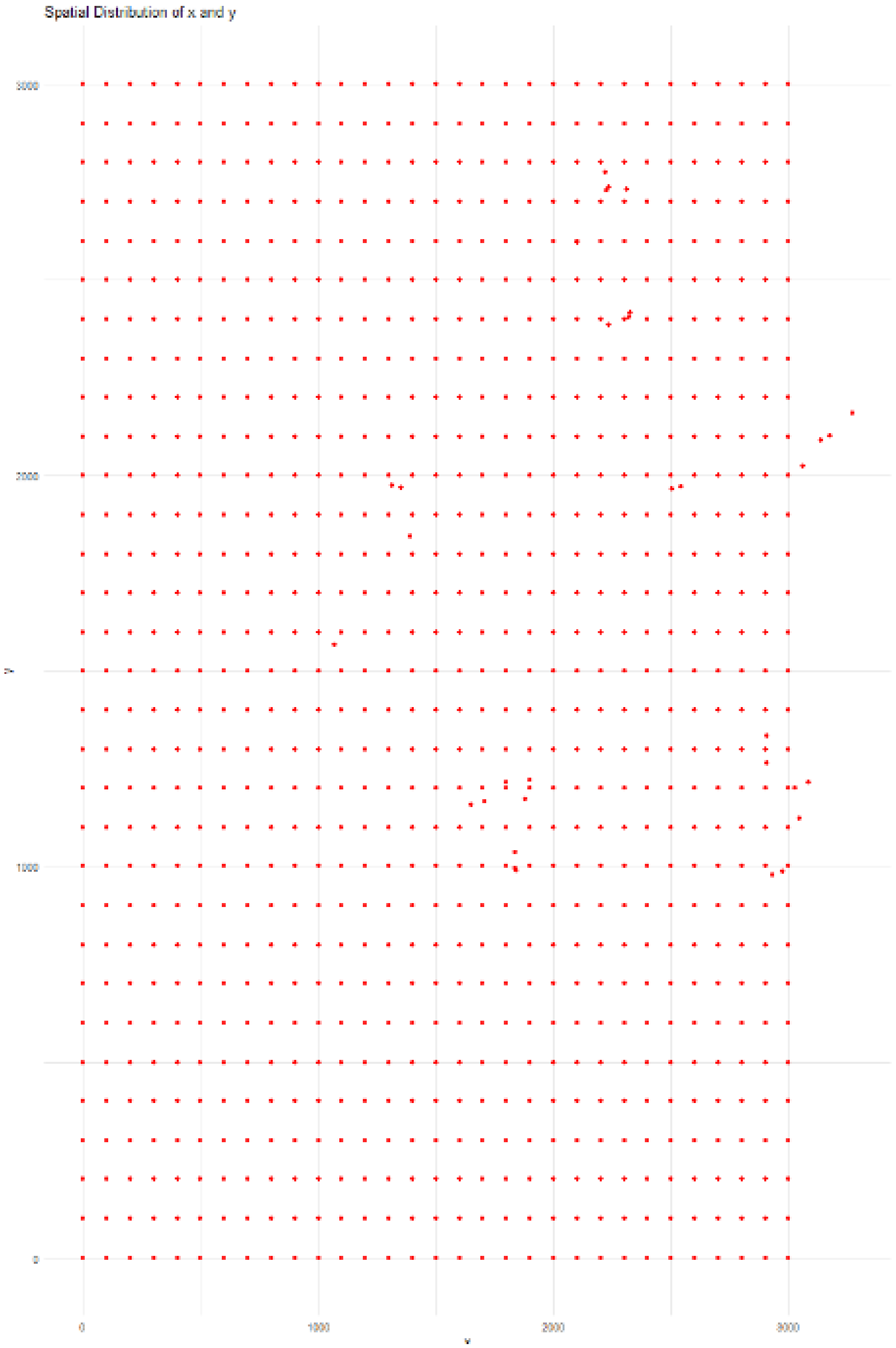

Figure 5.1, Spatial Distribution of *x* and *y*

6. **Which dependencies can you see?**

The correlation heatmap reveals a moderate positive correlation between *Co* and *Ni*, with a correlation coefficient of 0.52, indicating that regions with higher cobalt concentrations tend to have higher nickel concentrations. The scatterplot further confirms this trend, showing a clear positive relationship between these two elements.

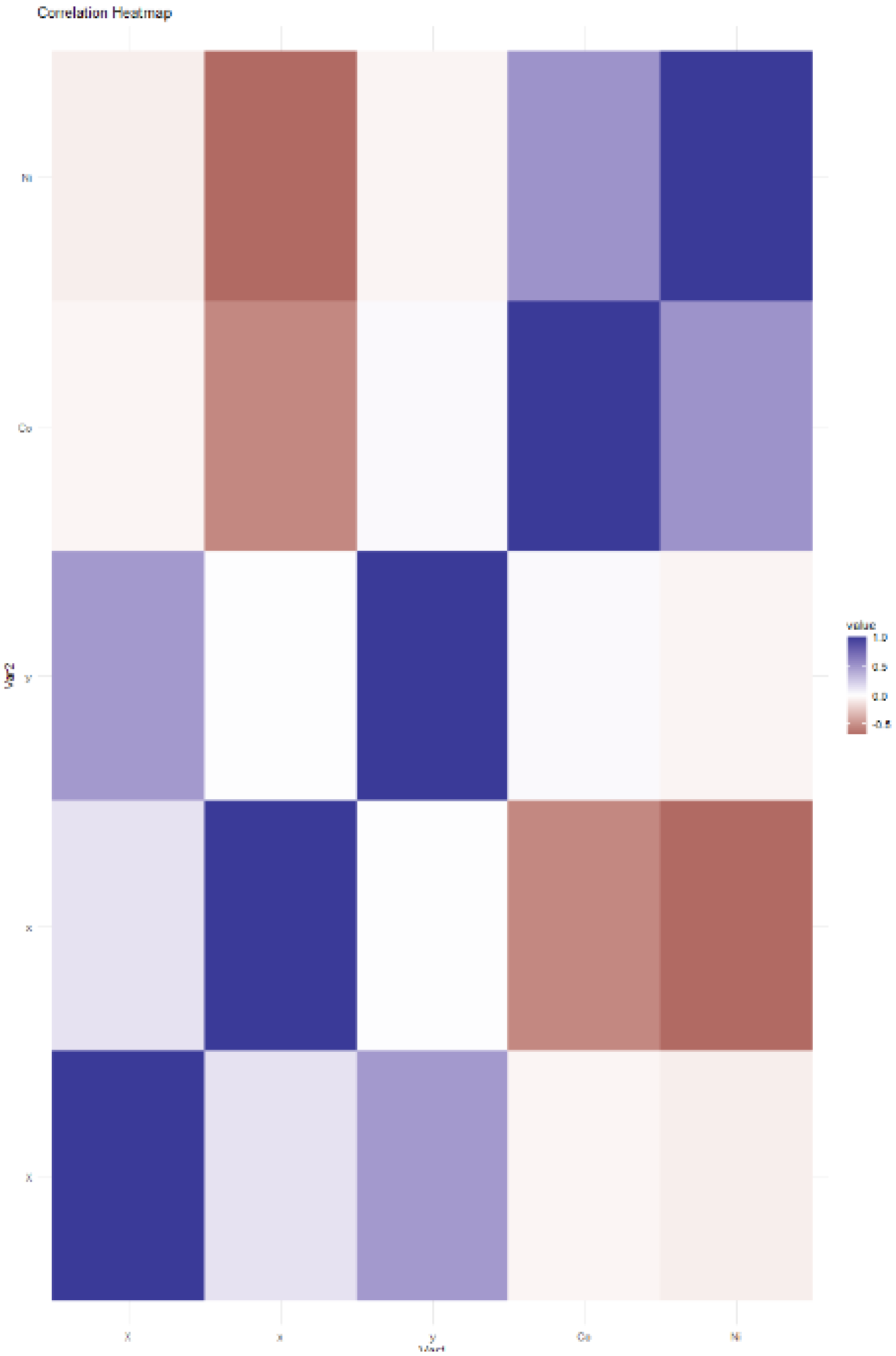

Figure 6.1, Correlation Heatmap

7. Are the dependencies strong or weak, linear or nonlinear, increasing or decreasing?

The correlation between *Co* and *Ni* is moderate and linear, as seen in the scatterplot and confirmed by the positive correlation value of 0.52. On the other hand, the relationship between *x* and *Ni* shows a weak, negative trend, suggesting that as the *x* value increases, the concentration of nickel tends to decrease, although the relationship is not strong enough to be considered a clear dependency.

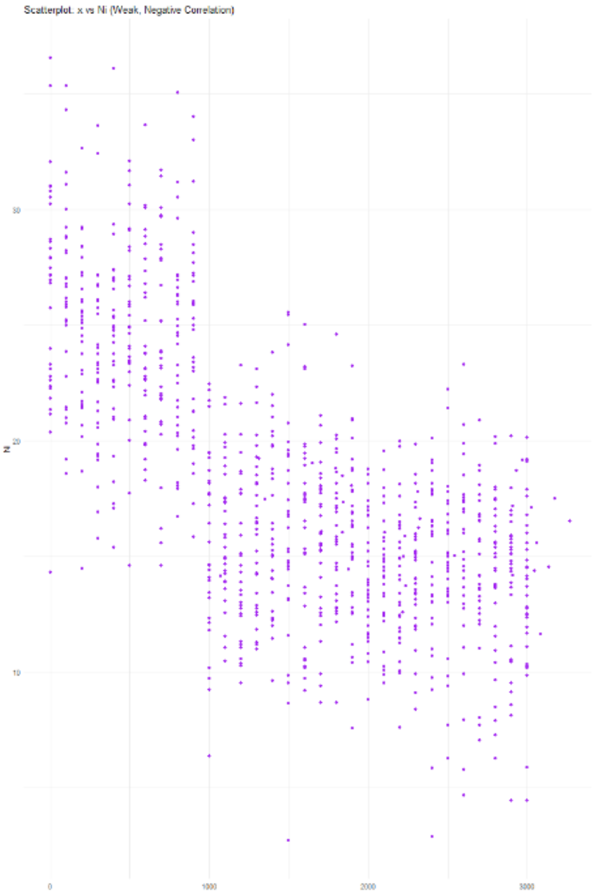

Figure 7.1 Scatterplot: x vs Ni (Weak, Negative Correlation)

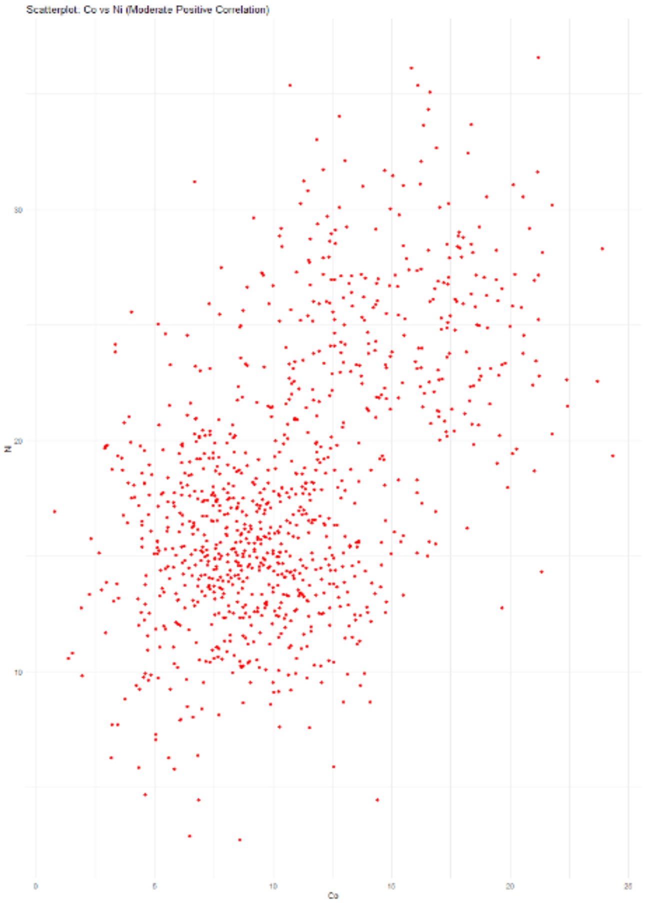

Figure 7.2, Scatterplot: Co vs Ni (Moderate Positive Correlation)

8. Do the observations correspond to what we would expect?

The scatterplot of y versus *Ni* shows no clear relationship, supporting the expectation that *y* might not influence the *Ni* concentration. This confirms that the column *y* is independent of *Ni*, as expected based on the earlier analysis that showed no correlation between y and other variables.

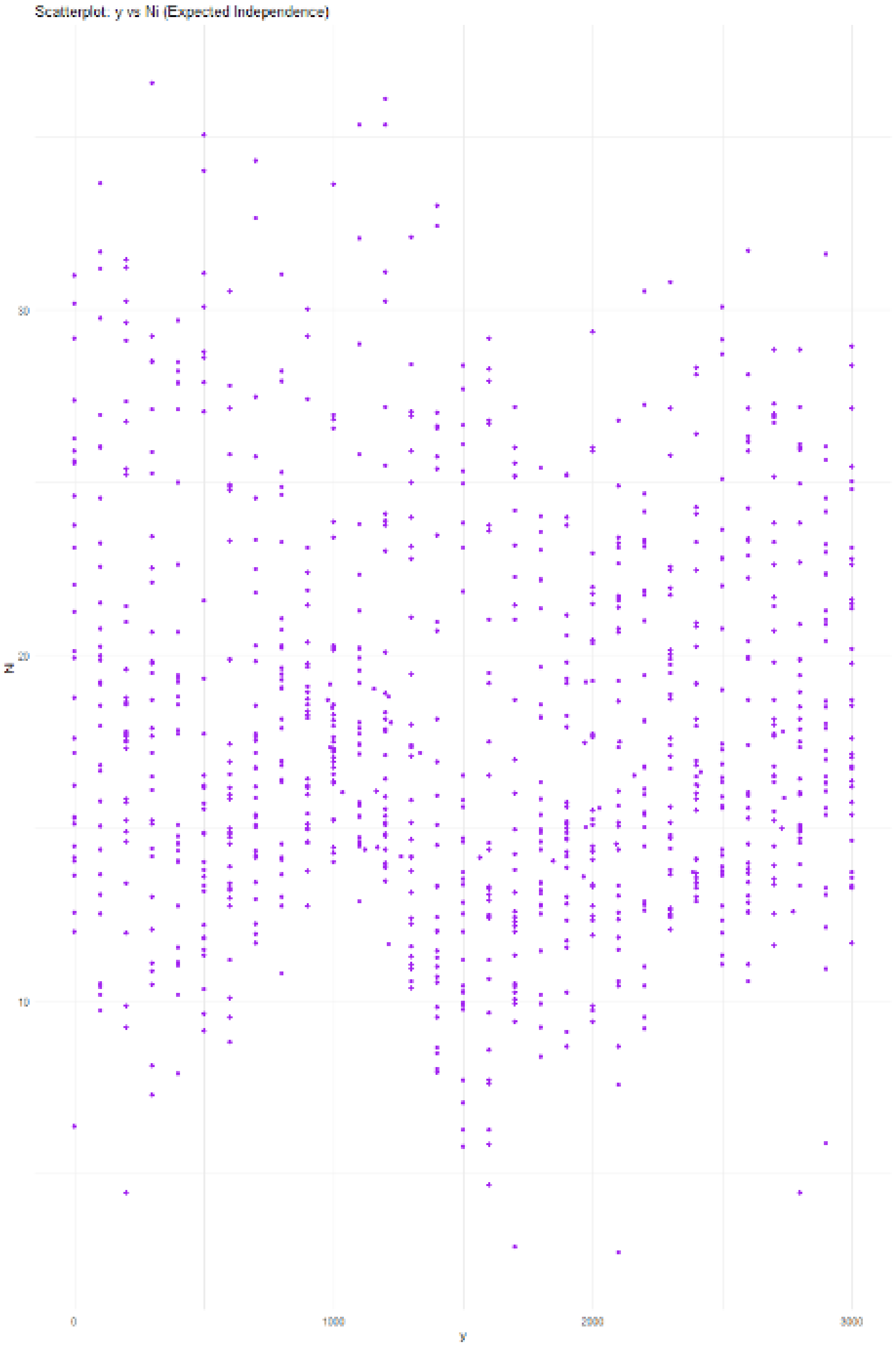

Figure 8.1, Scatterplot: y vs Ni (Expected Independence)

9. **Any other noticeable problems?**

The dataset does not contain any missing values, as confirmed earlier, but it's important to note that the *x* and *y* columns have limited unique values, which may indicate rounding or discretization. Furthermore, outliers in *Co* and *Ni* may require further investigation to determine whether they represent legitimate extreme values or data errors.

Missing Data Map
994
979
964
949
934
919
904
889
874
859
844
829
814
799
784
769
754
739
724
709
694
679
664
649
634
619
604
589
574
559
544
529
514
499
484
469
454
439
424
409
394
379
364
349
334
319
304
289
274
259
244
229
214
199
184
169
154
139
124
109
94
79
64
49
34
19
4
Ni
Co
y
x
Missing (0%)
Observed (100%)

Figure 9.1, Missing Data Map

Conclusion

The exploratory data analysis (EDA) of the dataset reveals several key insights into its structure and characteristics. The distributions of *x* and *y* are unusual, with a limited number of unique values, indicating that they likely represent structured grid-like data or coordinates, which may have been discretized or rounded. The variables *Co* and *Ni* show right-skewed distributions, with a concentration of values in the lower range and a few outliers suggesting potential extreme observations or data errors. Boxplots confirm the presence of outliers in both *Co* and *Ni*. The correlation analysis highlights a moderate positive correlation between *Co* and *Ni*, suggesting that higher concentrations of cobalt are associated with higher levels of nickel in the data. In contrast, the relationships between x and *Co* or *Ni* are weak and negative, indicating that as the spatial coordinate x increases, the concentrations of these elements tend to decrease slightly. Scatterplots support these findings by visually confirming the moderate, positive relationship between *Co* and *Ni* and the weak, negative trend between *x* and *Ni*. Additionally, the absence of a clear relationship between *y* and other variables further suggests that *y* may represent an independent or non-influential factor. The overall data quality appears to be high, with no missing values detected, although further investigation into the outliers and the spatial distribution might be necessary. These findings provide a comprehensive understanding of the dataset and lay the foundation for further analyses, such as predictive modeling or spatial trend analysis.