Assignment 3 – Data Analysis

Course Instructor: Nauman Ghazi

Date: 2022-03-06

Student Name	P.No	Contribution in the assignment (25 each % for equal contribution)
Mohit Battu	991007-T175	***25%***
Sai Chetan Poluri	*980501-0213*	***25%***
Monica Gattupalli	*991130-T308*	***25%***
Gowtham Kumar Sandaka	*001117-T071*	***25%***

Scenario-1

Question-1: What are the objects, subject, treatment, and factors for this experiment?

Answer:

Below are the object, subject, treatment, and factors according to the given experiment

	Subjects	**Objects**	**Treatment**	**Factors**
Design-A	Students	IDE-A, IDE-B	6 students work on IDE-A with prog-1 and 6 Students work on IDE-B with prog-1	Prog-1
Design-B	Students	IDE-A, IDE-B	All students work on both IDEs with the same prog 1	Prog-2
Design-C	Students	IDE-A, IDE-B	All the students work with 2 IDEs and with both progs.	Prog-1, Prog-2
Design-D	Students	IDE-A, IDE-B	All the students are divided into 4 groups, and they are assigned in a different order.	Prog-1, Prog-2

Assignment 3 – Data Analysis

Question-2: How would you descried Design-A and Design -D in terms of standard design type?

Answer:

According to the standard design type Design-A has 2 treatments and 1 factor and Design-D has 2 treatments and 2 factors

	Treatments	**Factors**	**Use**
Design-A	IDE-A, IDE-B	Prog-1	Students are equally divided to factor with the Treatment.
Design-D	IDE-A, IDE-B	Prog-1, Prog-2	Students should use the treatment to do the factors.

Question-3: What are the benefits and limitations of using Design-B instead of Design-A?

Answer:

Advantages of utilizing Design-B rather than Design-A

1. In Design -B, all students must complete the identical job in two IDEs so that we may compare the two IDEs' performance in executing Prog-1.

2. Allows you to justify the most flexible and efficient IDE for completing a specific activity under similar situations.

3. Each student will get the opportunity to learn how the two IDEs work.

4. We can also discuss how IDEs respond to specific programming scenarios.

The use of Design-B instead of Design-A has some drawbacks.

1. Each student must run the same Prog-1 program in two IDEs, which is an expensive and time-consuming operation.
2. The student may lose interest in continuing to do the same thing.

Question-4: What problems/mistakes can you identify in Design-C?

Answer:

Design-C is simply an advancement of Design-B that avoids the limitations of Design-B; additionally, in Design-C, students must execute two different programs in the IDE's that are equally difficult, which is an efficient way to compare the processing speed, performance, flexibility, and response time of the two IDE's. We may compare IDE's performance under various settings and limits using this design, but there is one disadvantage: the students may feel a little burdened.

Assignment 3 – Data Analysis

Question:5 Does Design-D solve the problems you have identified in Design-B and Design-C?

Answer:

Yes, Design-D has solved the restrictions of Design-B and Design-c. In Design-D, we simply give a single program to one student, which can be completed in the assigned IDE, reducing the workload, time, and expense. Students will be motivated to complete a single activity. In Design-D, there is just one aspect that produces comprehensive results, and we may draw conclusions about the features of the two IDEs in different situations based on this.

Question:6 What are the benefits and limitations of the designs Design-A and Design-D?

Answer:

Benefits of Design-A

1. Analyzing a single Factor is easy that is understanding a single program and executing that with any errors is very easy, it will take less amount of time in analyzing the program.
2. We can get the exact result as Design-A are dealing with a single program.

Benefits of Design-D

1. This Design is cost-effective and can be done with less amount of time, as each student needs to tackle with a single program in a single IDE.

Limitations of Design-A

1. In this design the students' experience and the knowledge may affect the result, because if the student has previous experience, then those students will do the problem in no time.

Limitations of Design-D

1. It is a little difficult to analyze 2 factors, that is understanding two programs and performing the task without any error is a little burden and difficult task to the students because all the students will not have the same IQ in performing the program.
2. In Design-D, it will have different combinations of students, program and the IDE in this case analyzing the outcome may be difficult because we may get a different review from each student which increases the time of concluding eth result.

Question7: What variables must be controlled in Design-A to increase the validity of the experiment?

Answer:

In Design-A, variables such as knowledge, expertise with that issue, grasping power, and students' IQ should be controlled. The results may be skewed due to the pupils' varying IQs and practical experience. If all the students are on the same level, the results will be more accurate. To improve the validity of the results, all students should have the same level of experience and should be familiar with utilizing IDEs.

Assignment 3 – Data Analysis

Answer the following questions related to analysis of an experiment with Design-A (as shown in Table 1) and the results in Table 5:

1. State the null and alternative hypothesis for this investigation.

Answer:

Null and Alternative hypotheses:

- ***Null hypotheses:***
  Programmers benefit from both IDEs, but none is superior to the other. They both perform similarly.
  H_0: {The required time to carry out IDE-A} = {The required time to carry out IDE-B}
- ***Alternative hypotheses:***
  Both IDEs make programming simpler and more efficient, but one performs better than the other.
  H_1: {The required time to carry out IDE-A > The required time to carry out IDE-B
  (OR)
  The required time to carry out IDE-A < The required time to carry out IDE-B}

2. Use descriptive statistics and visualize the data in Table 5 use e.g., box plot, histograms, and scatter plot. Which visualization tool helped you develop some insights into the data? What were the insights e.g., any interesting patterns or trends in the data, a clear difference in efficiency between two IDEs, outliers?

Answer:

In *Figure 1*, the descriptive statistics from Table 5 are shown. The total number of students involved in the experiment is 12. The presented scenario Table 5 shows the results of the Design-A experiment, in which half of the students used IDE-A to complete the task (i.e., program 1) and the other half used IDE-B to complete the task (i.e., program 1).

IDE-A

Student	Time (in minutes) to implement Prog-1
1	104
2	102
4	168
7	111
9	137
12	149

IDE-B

Student	Time (in minutes) to implement Prog-1
3	159
5	150
6	151
8	105
10	124
11	154

IDE-A

	Student	Time (in minutes) to implement Prog-1
count	6.000000	6.000000
mean	5.833333	128.500000
std	4.262237	27.061042
min	1.000000	102.000000
25%	2.500000	105.750000
50%	5.500000	124.000000
75%	8.500000	146.000000
max	12.000000	168.000000

IDE-B

	Student	Time (in minutes) to implement Prog-1
count	6.000000	6.000000
mean	7.166667	140.500000
std	3.060501	21.248529
min	3.000000	105.000000
25%	5.250000	130.500000
50%	7.000000	150.500000
75%	9.500000	153.250000
max	11.000000	159.000000

Figure 1: Descriptive Statistics of Table 5

Assignment 3 – Data Analysis

We can observe that from *Figure 1* the students achieved the mean execution time of 128.5 minutes for implementing program 1 using IDE-A. Using the IDE-B, the other half of the students attained a mean execution time of 140.5 minutes for completing the same work. The performance of the two IDEs is further plotted with the scatter plot to better visualize and understand the statistical data presented in *Figure 1*.

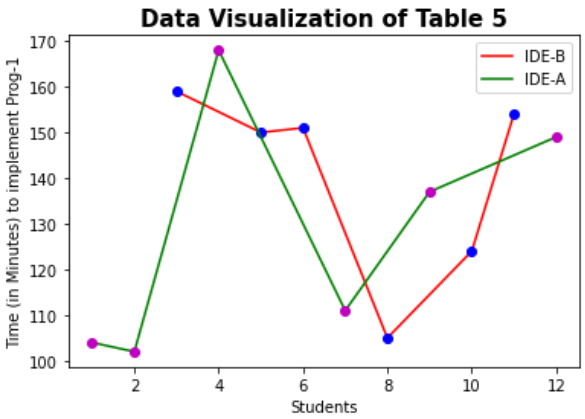


Figure 2: Scatter Plot Visualization of Table 5

Figure 2 shows a scatter plot depiction of the two groups of students who implemented program 1 using IDE-A and IDE-B. IDE-A is represented by the green line, whereas IDE-B is represented by the red line. *Figure 2* shows that the IDE-A had the longest execution time of 168 minutes and the shortest execution time of 102 minutes. Likewise, the IDE-B has attained a peak execution time of 159 minutes. When compared to the IDE-B, the above-scattered plot has clearly shown that the IDE-A takes less time to execute. Further to verify the data presented in two groups of students we have constructed the boxplots to ensure that there are no outliers in our given dataset.

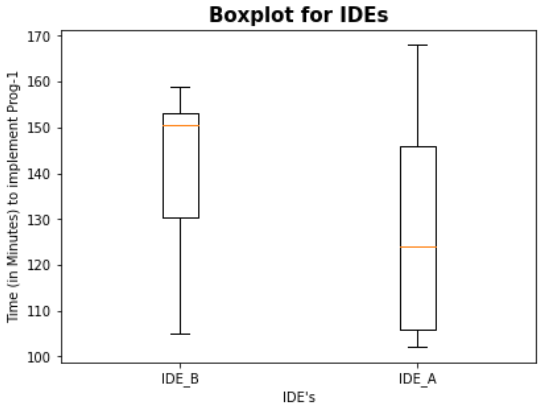


Figure 3: Boxplot Visualization of Table 5

Figure 3 shows the execution time of IDE-A and IDE-B in minutes. We may deduce from the Box plot graph that the data does not contain any outliers. In comparison to IDE-B, we can infer that IDE-A has achieved superior performance in terms of implementing program 1 in less amount of time.

3. Choose and justify your choice of a parametric/non-parametric test for analysing the given data (document the steps you undertook and the results).

- Given data is distributed data and mean can be considered as the most common measure used to find the center of the data that is distributed.

Assignment 3 – Data Analysis

- Parametric tests are usually used for distributed data and non-parametric tests are used for the Distribution – Free data.
- As we are going to find the mean of the distributed data, we can use the parametric tests to analyse the data.
- Initially we have considered data from table 1 and 5 from Design A from the given instructions document.
- Next the data is divided based on IDE, weather it is A or B.
- After analysing the data, we have tabulated the data along with the estimated time as tab 1 for IDE A and tab 2 for IDE B.

Student	Estimated time (in min)
1	104
2	102
4	168
7	111
9	137
12	149
Mean	**128.5**

Tab 1

Student	Estimated time (in min)
3	159
5	150
6	151
8	105
10	124
11	154
Mean	**140.5**

Tab 2

- Similarly, if we consider table 4 and 6 from the instruction set for Design D, the data is already categorised under IDE A and IDE B with a set of three students.
- We have again tabulated this as tab 3 tab 4, tab 5 and tab 6 along with the estimated time.

Student	Estimated time (in min)
1	71.3
2	110
9	115
Mean	**98.6**

Tab 3

Assignment 3 – Data Analysis

Student	Estimated time (in min)
3	178
7	94.9
12	109
Mean	**127.3**

Tab 4

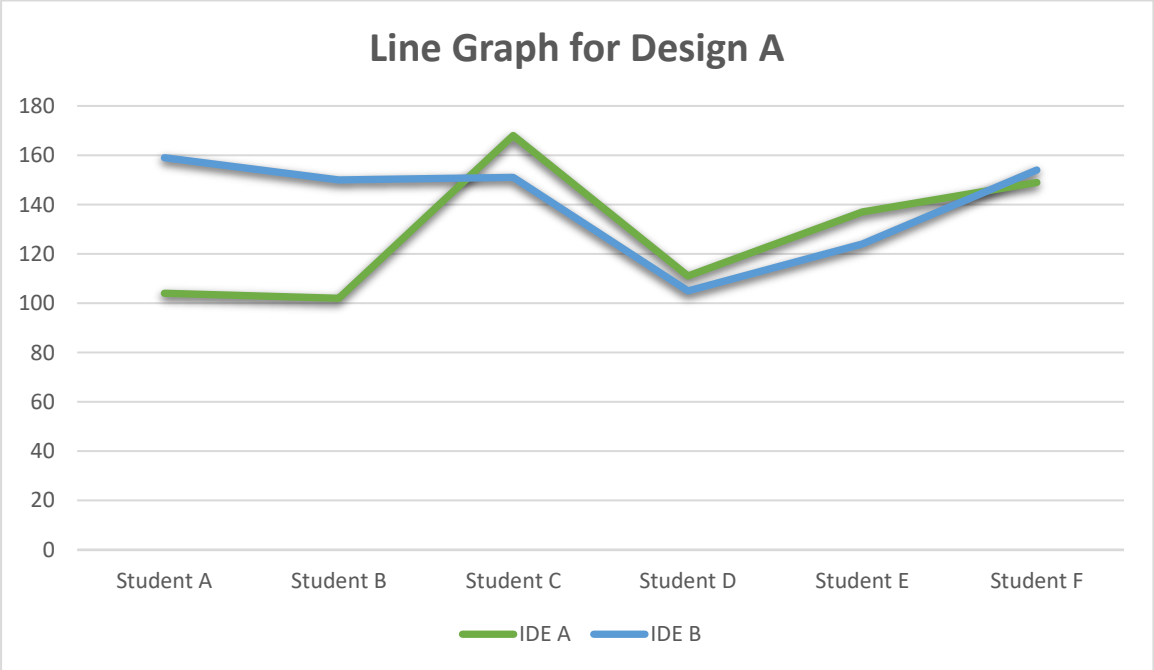
Student	Estimated time (in min)
4	153
6	174
11	95
Mean	**140.6**

Tab 5

Student	Estimated time (in min)
5	120
8	86.1
10	175
Mean	**127.03**

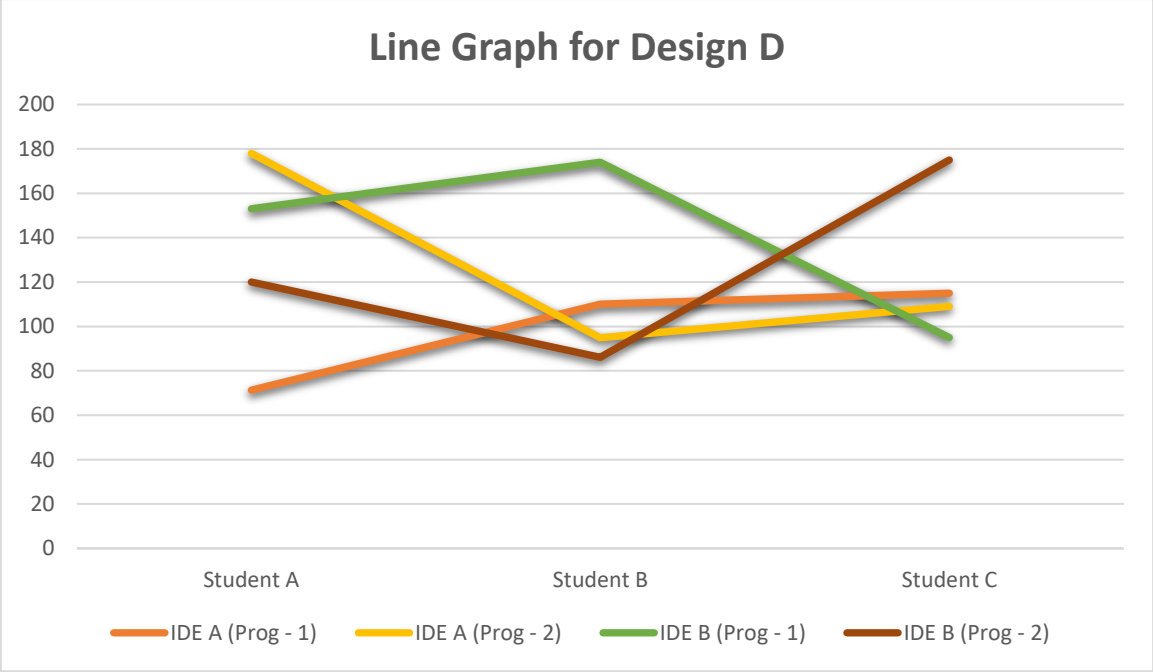
Tab 6

- It will be easy to understand if the data is plotted as graph.
- So, we have used this data and plotted a line graph.



Graph A

Assignment 3 – Data Analysis



Graph B

4. Run the statistical method and report if you can reject the null hypothesis? Please interpret your results, what does this imply for the objective of the study?

Answer:

We determined that IDE-A performed much better than IDE-B based on our observations and visualizations from Table 5. Let's verify if our claimed analysis is true or false using the statistical method approach.

The hypothesis assumed for the problem is stated below:

H_o: IDE-A = IDE-B (i.e., Null Hypothesis), where both the Integrated Development Environments are identical.

H_A: IDE-A ≠ IDE-B (i.e., Alternate Hypothesis), where we claim that IDE-A has achieved higher performance.

Given, the mean of IDE-A $\bar{x}_1$ is 128.50 (i.e., inferred from Figure 1).

The mean of IDE-B $\bar{x}_2$ is 140.50 (i.e., inferred from Figure 1).

Let's assume the α=0.05.

The standard deviation of IDE-A S_1 is 27.06 (i.e., inferred from Figure 1).

The standard deviation of IDE-B S_2 is 21.24 (i.e., inferred from Figure 1).

The population size of IDE-A and IDE-B are the same with a value of $n_1=n_2=6$.

The difference of the means for both the IDE-A and ID-B can be given as $\bar{x}_1 - \bar{x}_2 = 140.50-128.50 = -1$. Based on the sample size (i.e., 12) we have chosen the Unpaired t-test. We must first calculate the pooled Standard deviation on, then the Standard Error, and finally the t-statistics score.

The pooled standard deviation is then calculated by the below provided formula:

Assignment 3 – Data Analysis

$$S_p = \sqrt{\frac{(n_1 - 1) * (S_1)^2 + (n_2 - 1) * (S_2)^2}{n_1 + n_2 - 2}}$$

The n_1, n_2, S_1 and S_2 values are then substituted in the formula.

$$\sqrt{\frac{(6 - 1) * (27.06)^2 + (6 - 1) * (21.24)^2}{6 + 6 - 2}} = 24.32\ minutes$$

Therefore, the S_p value is 24.32 minutes.

The Standard Error is calculated by the below mentioned formula:

$$SE(\bar{x}_1 - \bar{x}_2) = S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$SE(\bar{x}_1 - \bar{x}_2) = 24.32 \sqrt{\frac{1}{6} + \frac{1}{6}} = 14.05\ minutes$$

Therefore, the $SE(\bar{x}_1 - \bar{x}_2)$ is 14.05 minutes.

The t-statistic score is calculated by the below mentioned formula:

$$T = \frac{\bar{x}_1 - \bar{x}_2}{SE(\bar{x}_1 - \bar{x}_2)}$$

$$T = \frac{-12}{14.05} = -0.85$$

Therefore, the T is -0.85.

The degree of freedom is further calculated by $DF = n_1 + n_2 - 2$

$$DF = 6 + 6 - 2 = 10$$

Therefore, the DF is 10.

From the provided data above, alpha is assumed to be 0.05 and the degree of freedom is 10, the critical value of the t-test is 1.812 (i.e., inferred from the T Table).

Since the critical value of t-test (i.e. 1.812) is greater than the score of the t-statistics (i.e. -0.85), we accept the null hypothesis that both the integrated development environments A and B are identical. As a result, both IDEs offer the same level of performance.

5. Based on the results would you be confident to recommend an IDE either IDE-A or IDE-B for use in your company. Why or why not?

Answer:

- We will go with the null hypothesis because the programs run by IDE-A and IDE-B are the same. Of course, the data visualization shows that IDE-A is more effective than IDE-B since IDE-A performed program-1 more effectively than IDE B. However, we

cannot suggest that our firm choose one IDE over the other. In terms of the hypothesis, it opposes the data visualization as the programming performance and outcomes are equal among IDEs. So, it'd be much better unless there were additional topics to put the IDE through its paces.

Scenario-2

a. Describe the approach that you will follow to analyze the given data (i.e., the three papers identified in Section 2.2). Please read Chapter 18 of C. Robson, K. McCartan, Real world research: A resource for social scientists and practitioner-researchers. Fourth Edition. Wiley, 2016, to make an informed decision about your approach and the steps you take. For example, the analysis approach you will use (a. Quasi-statistical approach, b. thematic coding approach, or c. grounded theory approach). Also, describe your mechanism for coding the data. Also, explain why you chose the approach over other alternatives.

Answer:

An analysis of the data provided was conducted using thematic coding. NVivo, a software application used for qualitative data analysis, was used to code the data. The analysis begins with an initial review of the data. Second, data coding is accomplished by clustering significant groups together. Additionally, groups are further clustered based on their detected themes.

A CAQDAS assisted us to establish a single storage area for many of the coded items; it allowed access to coded files quickly and easily and led to uniform coding.

Reasons for why we are using thematic coding:

- It is very adaptable and versatile, and it can be applied to a wide range of qualitative data.
- The results of theme coding analysis can be simply conveyed even when the other partner has no past knowledge of the topic.
- Thematic coding differs from other approaches to analysis in that it is simply understood and adjusted, allowing for analysis in a short amount of time.
- Fellow researchers have widely accepted the main approach to thematic analysis as a useful method of synthesizing significant elements from huge qualitative data sets.

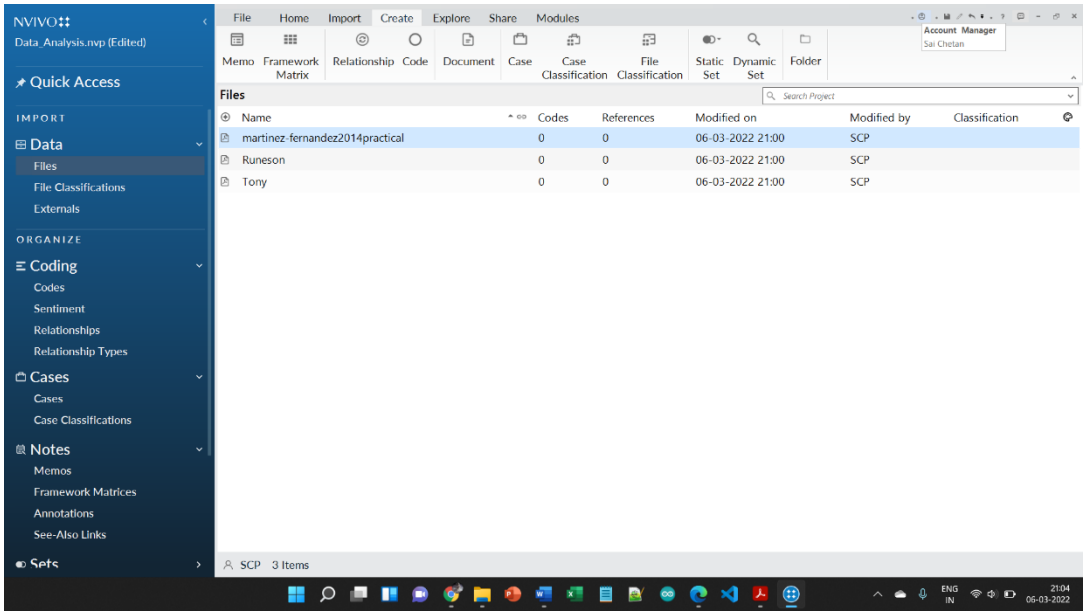
b. Please describe the coding procedure that you followed. For each step, please provide an example of how you coded the information in the papers (see Section 2.2 for the list of papers).

NVivo software is used to analyse the data given in the three papers.

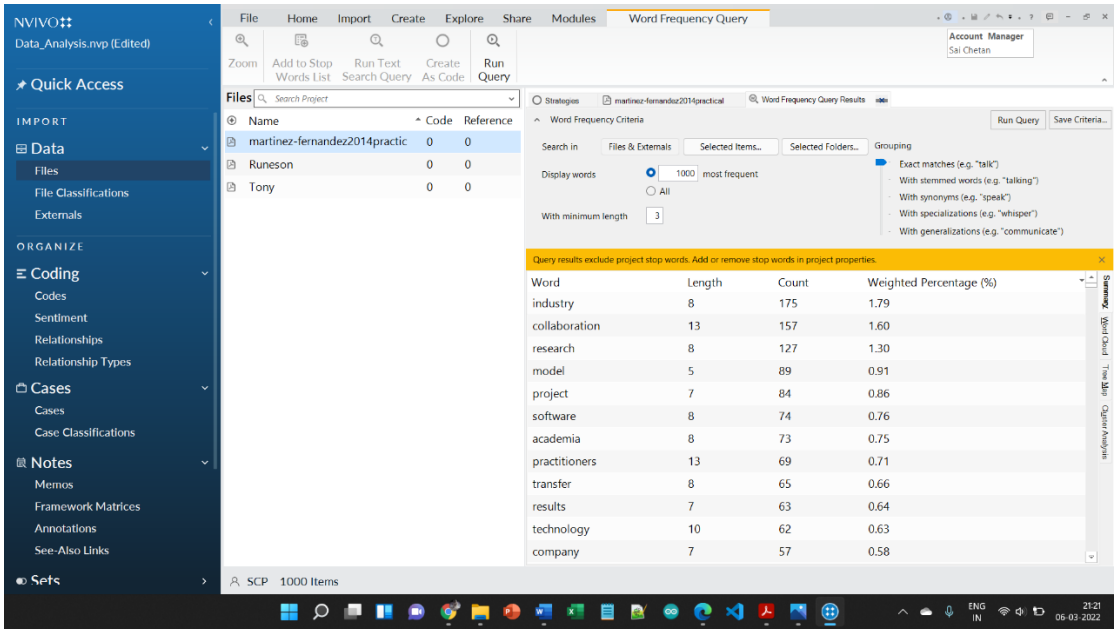
Coding Procedure:

- Initially we uploaded three papers into the NVivo Software using the import button in the software.

Assignment 3 – Data Analysis

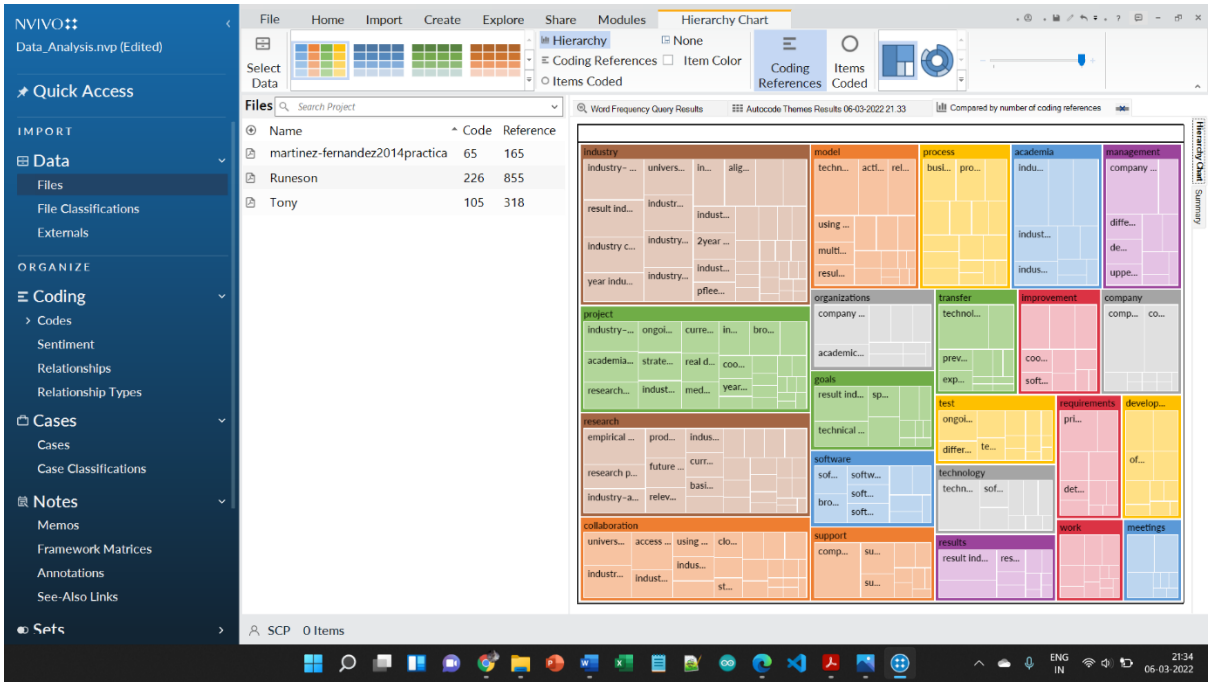


- As a second step we performed the query for the three imported files by displaying the words with a minimum length of three. Thus, the list of words is created.

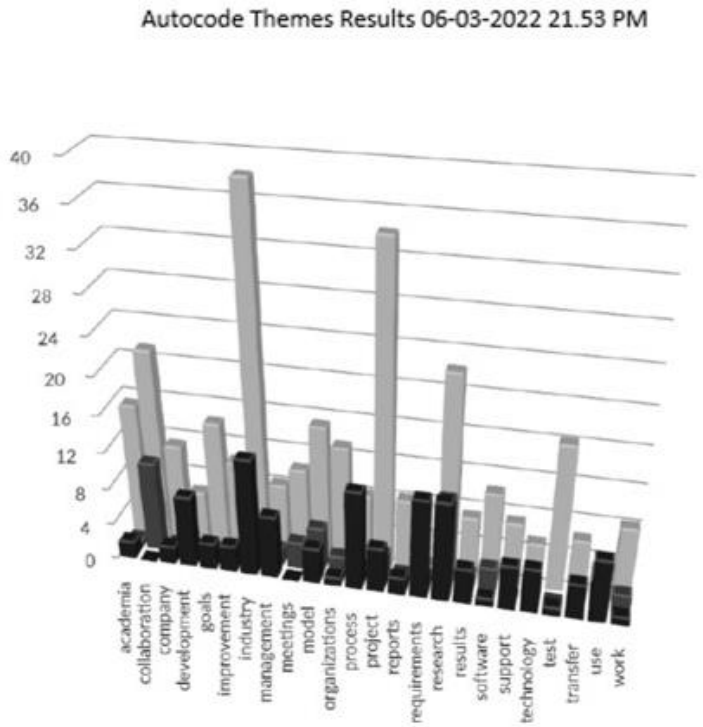


- Now we have used the Auto Code option which in turn identified the themes and generated the themes.
- Next Hierarchy Chars are used to represent these detected themes.

Assignment 3 – Data Analysis



- Finally, a 3-dimensional graph is plotted, which is given below.



Assignment 3 – Data Analysis

c. Answer the following questions by citing examples from your analysis of the three studies (provided in Section 2.2):

1. ***Which challenges or impediments for industry-academia collaborations have been raised by the papers?***

Answer:

- Collaboration team colleagues changed constantly as a result of regulations requiring workers to rotate after a certain amount of time.
- In order to be able to contribute to collaboration, practitioners who join it temporarily must have sufficient expertise.
- Researchers were unable to reach practitioners due to a lack of cooperation time.
- Absence of particular data because of confidentiality concerns and regulatory barriers to collaboration.
- It needs to be planned how the company and academia will pull and push each other in [1].
- Coordinating the time between industry and academia in [2].
- There seems to be a breakdown in interaction, development, and
- management among the researchers in [3].

2. ***What patterns have been proposed for industry-academia collaborations?***

Answer:

- Regular meetings to discuss will assist the researchers in interpreting current knowledge and gaining new insights.
- Group discussion allows for controlled conversation among collaborative team members.
- In order to continue to provide the required resources for the next step in the process, results should be communicated to higher management frequently.
- Whether highly sensitive data is necessary for the collaboration or not, the outcomes must be correct if the data is provided.
- It is recommended that the researcher devote his or her entire time to the collaborative practice and refrain from engaging in other activities.
- The models recommended for industry-academia interactions are shown in the figure below.

Assignment 3 – Data Analysis

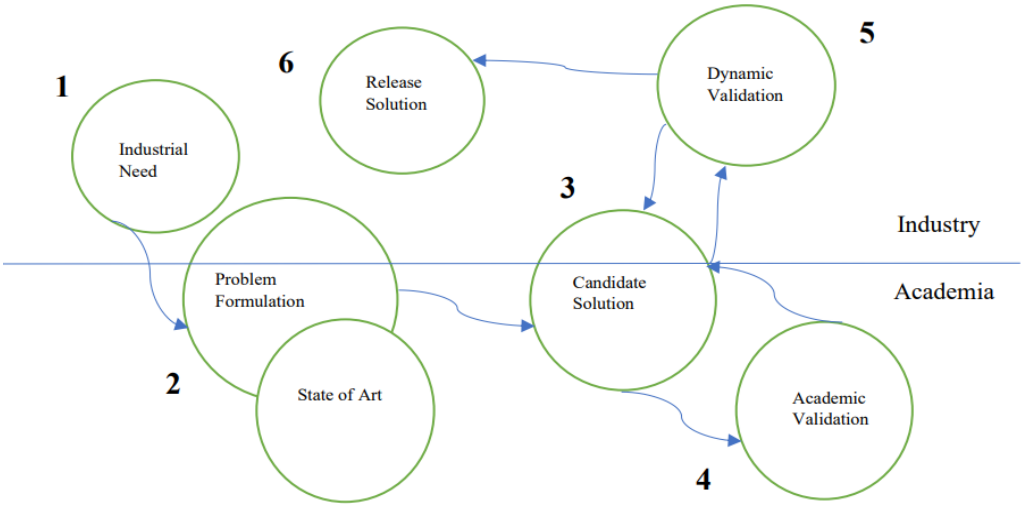


Figure 4: Industry-academia collaborations interactions

3. What should be avoided during industry-academia collaborations?

Answer:

- **Delay:** When it comes to industry-academia collaborations, it is indeed desirable to avoid delays whenever possible. It is imperative that all reports are sent on time; lags in collaboration aren't allowed.
- **Risks:** Risks can always be completely ignored, although they can be repurposed.
- **Low level of maturity:** A low maturity level requires more effective and compassionate handling of challenges. Because this is the key issue in industry-academia collaboration, it must be prevented throughout.
- **Communication:** The findings outcomes should have been adequately presented thus they'd be implemented in the industry.
- Communication breakdowns and a lack of confidence should be prevented.

References:

[1] P. Runeson, "It Takes Two to Tango–An Experience Report on Industry-Academia Collaboration," in *2012 IEEE Fifth International Conference on Software Testing, Verification, and Validation*, 2012, pp. 872–877.

[2] S. Martínez-Fernández and H. M. Marques, "Practical experiences in designing and conducting empirical studies in industry-academia collaboration," in *Proceedings of the 2nd International Workshop on Conducting Empirical Studies in Industry*, 2014, pp. 15–20.

[3] T. Gorschek, P. Garre, S. Larsson, and C. Wohlin, "A model for technology transfer in practice," *IEEE Softw.*, vol. 23, no. 6, pp. 88–95, 2006.