# OVERALL PLATFORM EVALUATION

## D24

IDEALVis Consortium

http://idealvis.inspirecenter.org/

European Union
European Regional Development Fund

Republic of Cyprus

Structural Funds
of the European Union in Cyprus

ΙΔΡΥΜΑ ΕΡΕΥΝΑΣ ΚΑΙ ΚΑΙΝΟΤΟΜΙΑΣ

# Executive Summary

This deliverable presents the results of the overall IDEALVis platform evaluation (Task 8.4). As described in Task 8.1, the primary aim of this evaluation was to assess the impact of the adaptation process on the user's performance and accuracy when interpreting data visualizations. Additionally, it aimed to assess the user acceptance and satisfaction with the platform.

To accomplish this, the present deliverable starts with the definition of appropriate acceptance and satisfaction metrics that were incorporated in the evaluation analysis. The key metrics selected for evaluating the platform were: (i) task performance; (ii) task accuracy; (iii) user experience; and (iv) platform usability.

Moreover, during the pilot study, each of the above metrics were recorded from the study participants at two instances. The first instance was when the participants were addressing analysis tasks using non-adapted/personalized data visualizations, and the second instance was when the participants engaged with data analysis tasks where the data visualizations were adapted/personalized according to each participant's unique user model.

The expected outcome for the successful (i.e., positive) platform evaluation is based on whether the metrics were positively influenced when participants interacted with the adaptative data visualizations; assuming that the baseline values for the metrics were the scores acquired from participants when using non adaptive data visualizations. The overall evaluation analysis of the platform revealed that the user experience and system usability factors were positively influenced by data visualization adaptation. Moreover, the performance and accuracy of participants were also positively influenced by data visualization adaptation, across specific data analysis task types, which we further explore in this deliverable.

The IDEALVis platform was evaluated, and the delivered adaptation was found to be effective in improving the user's (i) performance (i.e., time taken to address an analysis task) and accuracy (i.e., correctness of analysis task response), as well as (ii) the perceived user experience and platform usability scores. Regarding performance, users were faster by an average of 8.1 seconds when adaptation was enabled. Moreover, analysis task accuracy scores revealed that 62% of users were more accurate when responding to analysis tasks for which adaptation was enabled. Finally, adaptation impacted the users' perceived user experience score with an increase of 9%, and the reported platform usability score with an increase of 1.8%.

# Table of Contents

# List of Figures

# List of Tables

# 1 Introduction

This deliverable aims to evaluate the IDEALVis platform, according to data collected as part of the pilot study that took place towards the end of the project. The pilot study was conducted for the purpose of collecting data for the evaluation and assessment of the platform's functionalities regarding user requirements, user acceptance, satisfaction, performance, and accuracy. During the pilot study, the team was able to collect six datasets of information, which we utilised for the evaluation presented in this deliverable. Those datasets include user experience scores, system usability scores, and performance and accuracy scores (all scores were captured twice, i.e., before and after data visualization adaptation was provided). In Section 2 of this deliverable we present the goals of this evaluation, and based on these goals, we describe the underlying acceptance and satisfaction metrics that were incorporated into the evaluation analysis. Additionally, each of the acceptance and satisfaction metrics are ranked based on their importance and accountability regarding the overall platform evaluation. Moving on, in Section 3 we present the analysis of the platform evaluation results with regards to the participants' overall performance and accuracy when switching from the original non-personalized data visualization content to adapted content, which includes dynamically adapted/personalized data visualizations. Next, in Section 4 we provide the evaluation analysis results of the platform's user experience and usability factors. These factors were measured using a set of accredited system evaluation questionnaires before and after participants had received the adapted data visualization content, and we discuss the impact/effect of adaptation on those factors. Moreover, in Section 5 we provide an overall discussion with regards to the pilot study findings and possible next steps, and in Section 6 we present the conclusion of this deliverable.

# 2 Acceptance and Satisfaction Metrics

The overarching goal of IDEALVis, since its inception, was to further advance the data analytics field by enabling human-centred adaptive data visualizations in the context of the business domain. Taking multiple steps towards this goal, the project initially defined a novel multi-dimensional human-centred user model that incorporates cognitive factors, domain expertise and experience. Additionally, through a set of experiments and user studies, the project examined the impact of several human factors on data visualizations, and using this knowledge, formulated a set of adaptation rules for defining an adaptive data visualizations framework. The goal of this framework was to leverage both, the user model and the adaptation rules, for recommending the most appropriate data visualization for the unique end-user by altering the data visualization's type, structure, and semantics. While the main goal was the delivery of personalised data visualizations, another objective of the project was to further encapsulate the innovative data visualization framework and its surrounding components in an intuitive data analysis system i.e., the IDEALVis platform.

## 2.1 Evaluation Metrics

While the introduction of this section discussed the goals of the project and the IDEALVis platform, it was primarily focused on the project's expected outcomes regarding the actual technical developments. In this subsection we discuss the metrics that were used to evaluate the IDEALVis platform and its adaptation components, in terms of achieving its primary aims and goals, which were to facilitate more efficient and effective data exploration, thus enabling more effective decision making on critical business tasks. The list of acceptance and satisfaction metrics that were used are presented in order of importance (high importance first) in Table 1. This set of metrics are the key indicators used in the analysis presented throughout the next sections of this deliverable.

*Table 1 - Acceptance and Satisfaction Metrics*

| Importance Priority | Metric | Description |
|---|---|---|
| 1 | Increased Analysis Task Performance in milliseconds (ms) | Measures the average performance gain (expressed as decrease in milliseconds) exhibited by participants when working on adapted data visualizations. Performance achieved with non-adapted/non-personalised data visualizations is used as a baseline. |
| 2 | Increased Analysis Task Accuracy | Measures the increase in accuracy (expressed as total number of tasks addressed correctly) exhibited by participants when working on adapted data visualizations. Accuracy achieved with non-adapted/non-personalised data visualizations is used as a baseline. |
| 3 | Increase in Pragmatic Quality | Measures the increase of pragmatic quality scores (one of two user experience metrics) exhibited by participants when working on adapted data visualizations. Pragmatic quality scores achieved |

| | | |
|---|---|---|
| | | with non-adapted/non-personalised data visualizations are used as a baseline. |
| 4 | Increase in Usability Score | Measures the increase of usability scores (system usability metric) exhibited by participants when working on adapted data visualizations. Usability scores achieved with non-adapted/non-personalised data visualizations are used as a baseline. |
| 5 | Increase in Hedonic Quality | Measures the increase of hedonic quality scores (one of two user experience metrics) exhibited by participants when working on adapted data visualizations. Hedonic quality scores achieved with non-adapted/non-personalised data visualizations are used as a baseline. |

# 3 Adaptation's Impact on Accuracy and Performance

In this section we explore the analysis results of the IDEALVis platform evaluation, in terms of its ability to enable more efficient and effective data analysis and exploration through data visualization adaptation. Prior to presenting the evaluation results, we describe the overall pilot study, including its goals, setup, design, and materials used for successful execution. Moreover, we discuss the study procedure, including the training and onboarding phases performed for existing but also new participants. Finally, we present the results of the analysis performed on the evaluation metrics captured through the pilot study (i.e., data regarding the performance and accuracy of participants when interacting with various adapted and non-adapted data visualizations for addressing data analysis tasks), focusing on how IDEALVis adaptation mechanisms impacted the participants' overall performance and accuracy when addressing data analysis tasks.

## *3.1 The Pilot Study*

The pilot study was one of the most essential steps towards the successful evaluation and finalization of the IDEALVis project. Its primary goal was to implement the appropriate procedures that would drive the validation and assessment of the platform's functionalities regarding user requirements, user acceptance, satisfaction, and engagement. As mentioned in Section 2 of this deliverable, the overarching goal of this project is to enable data analysts to achieve a more efficient and effective data exploration of business datasets, using adaptive/personalised data visualizations as the primary means to achieve that. To this end, in order to quantitatively evaluate the IDEALVis platform's functionality and components (i.e., the adaptation and visualization engines, the set of adaptation rules, etc.), the pilot study was carefully designed to investigate the impact on the study participants' efficiency and effectiveness while utilizing the platform's components for data analysis. In the next sections, we discuss the setup, design, and materials used for this study, as well as the study procedure followed.

### 3.1.1 STUDY SETUP AND DESIGN

Since the primary goal of this study was to evaluate the effectiveness of the developed adaptation components in terms of enabling the participants achieve better efficiency and effectiveness when solving data analysis tasks, it was important to capture the participants' performance and accuracy when addressing data analysis tasks using: (a) the original non-adapted/personalized content, which includes data visualizations generated from the datasets without any alterations or enhancements (i.e., control tasks) and (b) the adapted content, which includes dynamically adapted/personalized data visualizations derived from applying the mapping rules and adaptive interventions, based on the participant's unique user model (i.e., personalized tasks).

Furthermore, for the purpose of the study we had to construct a realistic dataset along with two sets of matching visual analysis tasks (i.e., similar in terms of task type and complexity) that were based on the constructed dataset. The first set of analysis tasks was used in the first part of the study where the data visualization used to address each of the analysis tasks were not adapted i.e., the system returned the same data visualization for the specific analysis task, across all participants. The second set of analysis tasks was used in the second part of the study where the data visualizations used to address each of the analysis tasks were automatically adapted by the system, according to the individual performing the analysis.

### 3.1.2 STUDY MATERIALS

**Analysis Dataset:** The dataset used in the pilot study was constructed by the project team in accordance with the experience and expertise of the collaborator organizations from where the study participants were recruited. Moreover, those organizations offered sample datasets which the team transformed and prepared according to the goals of this study. The finalised dataset that emerged after the transformations was about Soft Drink Sales. The dataset is a transactional dataset composed of 731,446 observations. In Table 2 we list the attributes of this dataset along with their datatype and description. Moreover, it was decided that during the study, participants would be given the role of a Brand Manager working for a soft drinks company that has the product called IdealCola.

*Table 2 – Pilot Study Dataset's Attributes*

| Attribute Name | Data Type | Description |
|---|---|---|
| TDATE | Date | Full date of the transaction |
| YEAR | Integer | Represents the year of the transaction e.g., 2020 |
| MONTH | Integer | Represents the month of the transaction i.e., 1 to 12 |
| DAY | Integer | Represents the day of the transaction e.g., 1 to 30 |
| QUARTER | Integer | Represents the quarter of the transaction i.e., 1 to 4 |
| BRAND | Nominal | The name of the transaction's product brand |
| PRODUCT_NAME | Nominal | The transaction's product name |
| PROMOTION | Nominal | Promotion regarding this transaction. One of 5 promotion categories |
| PACK_TYPE | Nominal | Pack type regarding the product of the transaction. One of 3 pack type categories |
| DIET | Boolean | Whether or not the transaction's product is a diet product |
| OUTLET_NAME | Nominal | Name of the outlet where the transaction was made |
| OUTLET_TYPE_NAME | Nominal | Type of the outlet where the transaction was made. One of 9 outlet type categories |
| URBAN_RURAL | Nominal | Area of outlet where the transaction was made including if the area is urban or rural e.g., Famagusta Rural. |
| AREA_NAME | Nominal | Area of outlet where the transaction was made. One of 5 cities |
| M_SIZE | Float/Continues | Size of the transaction's product |
| M_PRICE | Float/Continues | Price of the transaction's product |
| M_QUANTITY | Float/Continues | The quantity of the product bought in this transaction e.g., IdealCola x2 |

| | | |
|---|---|---|
| M_SALES_VALUE | Float/Continues | M_QUANTITY multiplied by M_PRICE |
| M_SALES_VOLUME | Float/Continues | M_QUANTITY multiplied by M_SIZE |

**Analysis Tasks:** All analysis tasks for the study were built using the dataset mentioned above, while having in mind the participants' fictional role of a Brand Manager. The 39 analysis tasks that were constructed for this study can be seen in Table 3. The analysis tasks in Table 3 are split in 19 pairs of tasks, with each pair including (i) the control non-adapted/non-personalised task which is to be addressed by all participants using a specific/predefined visualization type, and (ii) a corresponding analysis task which will utilise the system's adaptation engine to return the adapted/best fit data visualization according to the participant addressing that particular task. Additionally, each pair of tasks has a specific analysis task type, which follows the same taxonomy used for building analysis tasks in our second user study mentioned in deliverable D11 - The Impact of Cognitive Factors on Data Visualizations. We followed the same taxonomy of data analysis tasks, since our adaptation rules were built using the participants' performance captured during the second user study and while they were interacting with this taxonomy of visualization tasks.

*Table 3 - Pilot Study Analysis Tasks*

| Task Name | Task Narrative | Task Type | Visualization Used |
|---|---|---|---|
| T01 Control Task | Identify the month with the highest sales during 2021 for brand "IdealCola". | Simple Comparison | Bar Chart |
| T01 Adaptive Task | Identify the month with the highest sales during 2021 for product "IdealCola Zero .33ltr x8 Can". | | Adapted Data Visualization |
| T02 Control Task | Identify the 3 top brands with the highest sales in 2019. | Retrieve Value | Data Table |
| T02 Adaptive Task | Identity the third (3) best-selling outlet type in 2020 in terms of sales volume. | | Adapted Data Visualization |
| T03 Control Task | Identity the second (2) best-selling area in 2021 for your brand IdealCola, in terms of sales value. | Retrieve Value | Pie Chart |
| T03 Adaptive Task | Identify if plastic or glass bottle is the third (3) best-selling pack type in 2021 for your brand IdealCola, in terms of sales value. | | Adapted Data Visualization |
| T04 Control Task | Identify if your brand IdealCola is growing during the first semester of 2021 (January to June) in terms of sales value. | Simple Comparison | Line Chart |
| T04 Adaptive Task | Identify if the glass bottles pack type is growing in terms of sales value in August of 2021 compared to June 2021. | | Adapted Data Visualization |

| | | | |
|---|---|---|---|
| T05 Control Task | Identify if the sales of IdealCola in Limassol are growing during the fall months of 2019 (September to November) in terms of sales value. | Simple Comparison | Line Chart |
| T05 Adaptive Task | Identify if the diet soft drinks sales are growing during the summer months of 2019 (June to August) in terms of sales value. | | Adapted Data Visualization |
| T06 Control Task | Identify if the sales of the Soft Drinks category are growing during the last 3 years in terms of sales value. | Simple Comparison | Column Chart |
| T06 Adaptive Task | Identify if the sales of the product "IdealCola .33ltr Can" are growing during the last 3 years (2019-2021) in terms of sales value. | | Adapted Data Visualization |
| T07 Control Task | Identify the volume of sales for Bakery outlet type during 2019 in thousands. For example, if you discover that the sales are 1,234,567 then you should report only the thousands' part i.e., 1234. You should not perform any rounding. | Retrieve Value | Column Chart |
| T07 Adaptive Task | Identify the volume of sales for Famagusta area during 2020 in thousands. For example, if you discover that the sales are 1,234,567 then you should report only the thousands' part i.e., 1234. You should not perform any rounding. | | Adapted Data Visualization |
| T08 Control Task | Identify the key competitor of brand LegendarySoda during 2020. The key competitor is not the top selling brand but the closest brand with higher sales than the brand in question. | Simple Comparison | Data Table |
| T08 Adaptive Task | Identify the key competitor of brand Crush during 2020. The key competitor is not the top selling brand but the closest brand with higher sales than the brand in question. | | Adapted Data Visualization |
| T09 Control Task | Identify which outlet type you will target to launch a new soft drinks product in 2022. You should opt for the outlet type that holds the majority of sales during the last year (i.e., 2021). | Simple Comparison | Pie Chart |
| T09 Adaptive Task | Identify which area you will target to launch a new soft drinks product in 2022. You should opt for the area that holds the majority of sales during the last year (i.e., 2021). | | Adapted Data Visualization |
| T10 Control Task | During 2021, between March and July, your IdealCola brand experienced consecutive rises in sales. At the same time the shares of IdealCola decreased. Identify the reason behind this. | Correlation | Bar |
| T10 Adaptive Task | During the fall of 2019, between September and November, your IdealCola brand experienced consecutive decreases in sales. At the same time | | Adapted Data Visualization |

| | | | |
|---|---|---|---|
| | there was a rise in overall shares. Identify the reason behind this. | | |
| T11 Control Task | Identify if the brand SteviaCola is affected by seasonality during the last 3 years (2019-2021). | Anomaly Detection | Line Chart |
| T11 Adaptive Task | Identify if the Soft Drinks category is affected by seasonality during the last 3 years (2019-2021). | | Adapted Data Visualization |
| T12 Control Task | Identify which month disrupts the pattern of monthly sales in 2021 for Hypermarkets. | Correlation | Column Chart |
| T12 Adaptive Task | Identify which area has a different trend compared to the other ones in terms of monthly sales in 2021. | | Adapted Data Visualization |
| T13 Control Task | Identify if promotions have a significant impact on overall sales in 2021 for all soft drinks. | Compute Derived Value | Column Chart |
| T13 Adaptive Task | Identify if the percentage of diet soft drinks' sales are between 40-50% of regular soft drinks' sales. | | Adapted Data Visualization |
| T14 Control Task | Identify in which month of 2021 did the sales of LegendarySoda outperform the sales of DreamSoda. | Compute Derived Value | Line Chart |
| T14 Adaptive Task | Identify the number of months where Convenience stores outperform the sales of bakeries for soft drinks during the last 3 years. | | Adapted Data Visualization |
| T15 Control Task | Identify your key competitor's (FizzySoda) strongest product (with regards to sales), during 2021. | Simple Comparison | Column Chart |
| T15 Adaptive Task | Identify the month with the lowest sales during the last quarter of 2020 for product "IdealCola Light .33ltr x8 Can". | | Adapted Data Visualization |
| T16 Control Task | Identity the second least selling outlet type in 2021 for your brand IdealCola, in terms of sales value. | Simple Comparison | Pie Chart |
| T16 Adaptive Task | Identify which quarter of 2021 your brand IdealCola had the highest sales. | | Adapted Data Visualization |
| T17 Control Task | You are currently distributing IdealCola to all districts. Identify which district you will avoid (according to the sales of 2021) to decrease your distribution cost. | Simple Comparison | Data Table |
| T17 Adaptive Task | You are currently distributing IdealCola to all outlet types. Identify which outlet type you will avoid (according to the sales of 2021) to decrease | | Adapted Data Visualization |

| | | | |
|---|---|---|---|
| | your distribution cost. Please specify only the first 4 letters of the outlet type in your answer. | | |
| T18 Control Task | Using all IdealCola brand past sales value data, identify the sales value of IdealCola brand for April 2022. In order to answer this question, you need to Forecast the sales value of IdealCola for 4 months. | Retrieve Value | Line Chart |
| T18 Adaptive Task | Using all past sales value data, identify the sales value of all soft drink brands for January 2022. In order to answer this question, you need to Forecast the sales value of all brands for a single month. | | Adapted Data Visualization |
| T19 Control Task | Using all Supermarket past sales value data, identify whether January or April 2022 will have the lowest value in sales for the Supermarket outlet type. In order to answer this question, you need to Forecast the sales value of all Supermarkets for 4 months. | Simple Comparison | Line Chart |
| T19 Adaptive Task | Using all diet soft drink past sales value data, identify whether June or August 2022 will have the highest value in sales for diet soft drinks. In order to answer this question, you need to Forecast the sales value of all diet brands for 8 months. | | Adapted Data Visualization |

**The Platform:** The platform received a couple of updates to accommodate the pilot study design. Those updates include

i. an administrative interface for adding, managing and assigning analysis tasks to analyst participants (Figure 1),

ii. a page where the participant can set a specific analysis task as the current task being addressed, including appropriate controls that enable the participant to provide a response to the current analysis task (Figure 2),

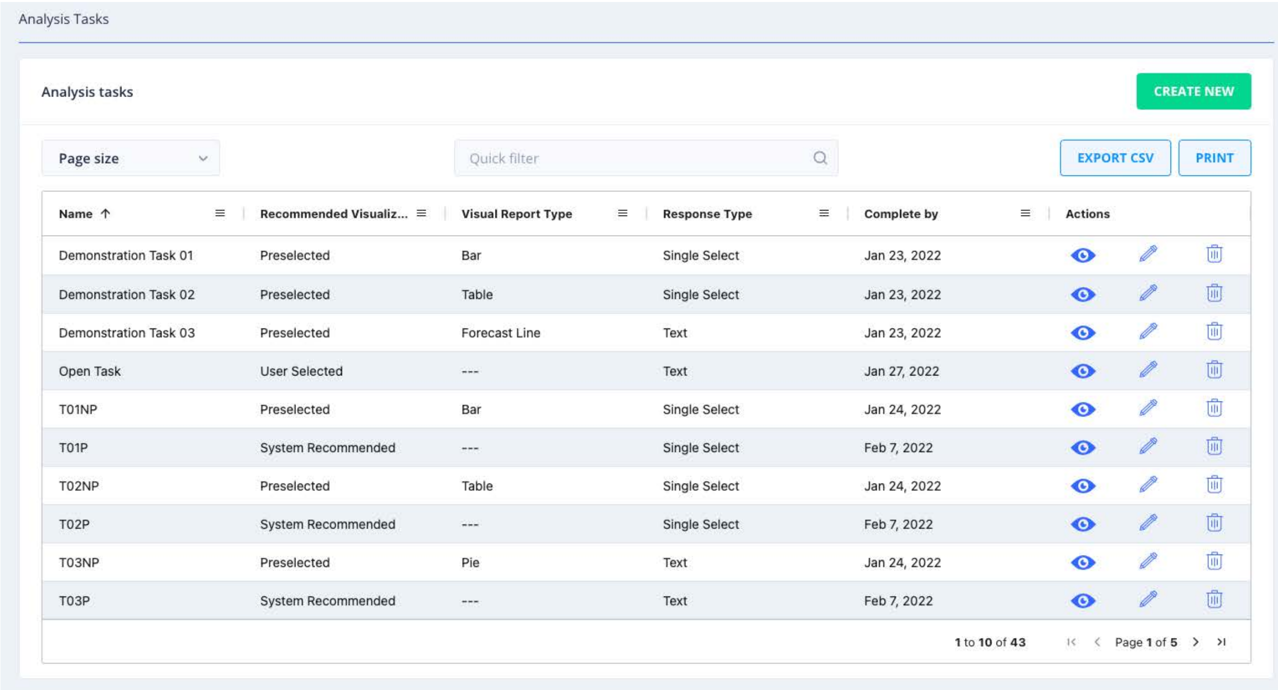
iii. a modal of analysis tasks instructions in the analysis wizard that guides the participant in how to perform the required analysis for the current analysis task (Figure 3),

iv. changes to the dashboard allowing the participant to pin a specific data visualization for the current task being addressed

v. minor changes to the tracker mechanism that is responsible for capturing the time a participant looks at a specific data visualization for a given analysis task (i.e., performance in milliseconds) and finally

vi. a new mechanism that resets the analysis task if the participant is found to be non-responsive for a number of seconds (Figure 4).

The last update (vi) was developed to increase the control of the study (since it was run remotely) and the quality of the data being collected, by mitigating the instances where a participant had a data visualization on screen while being away from the computer and thus floating the system with extremely high (in terms of milliseconds) tracking records.




| Name | Recommended Visualiz... | Visual Report Type | Response Type | Complete by | Actions |
|---|---|---|---|---|---|
| Demonstration Task 01 | Preselected | Bar | Single Select | Jan 23, 2022 | |
| Demonstration Task 02 | Preselected | Table | Single Select | Jan 23, 2022 | |
| Demonstration Task 03 | Preselected | Forecast Line | Text | Jan 23, 2022 | |
| Open Task | User Selected | --- | Text | Jan 27, 2022 | |
| T01NP | Preselected | Bar | Single Select | Jan 24, 2022 | |
| T01P | System Recommended | --- | Single Select | Feb 7, 2022 | |
| T02NP | Preselected | Table | Single Select | Jan 24, 2022 | |
| T02P | System Recommended | --- | Single Select | Feb 7, 2022 | |
| T03NP | Preselected | Pie | Text | Jan 24, 2022 | |
| T03P | System Recommended | --- | Text | Feb 7, 2022 | |

*Figure 1 - Administrative Interface for Handling Analysis Tasks*

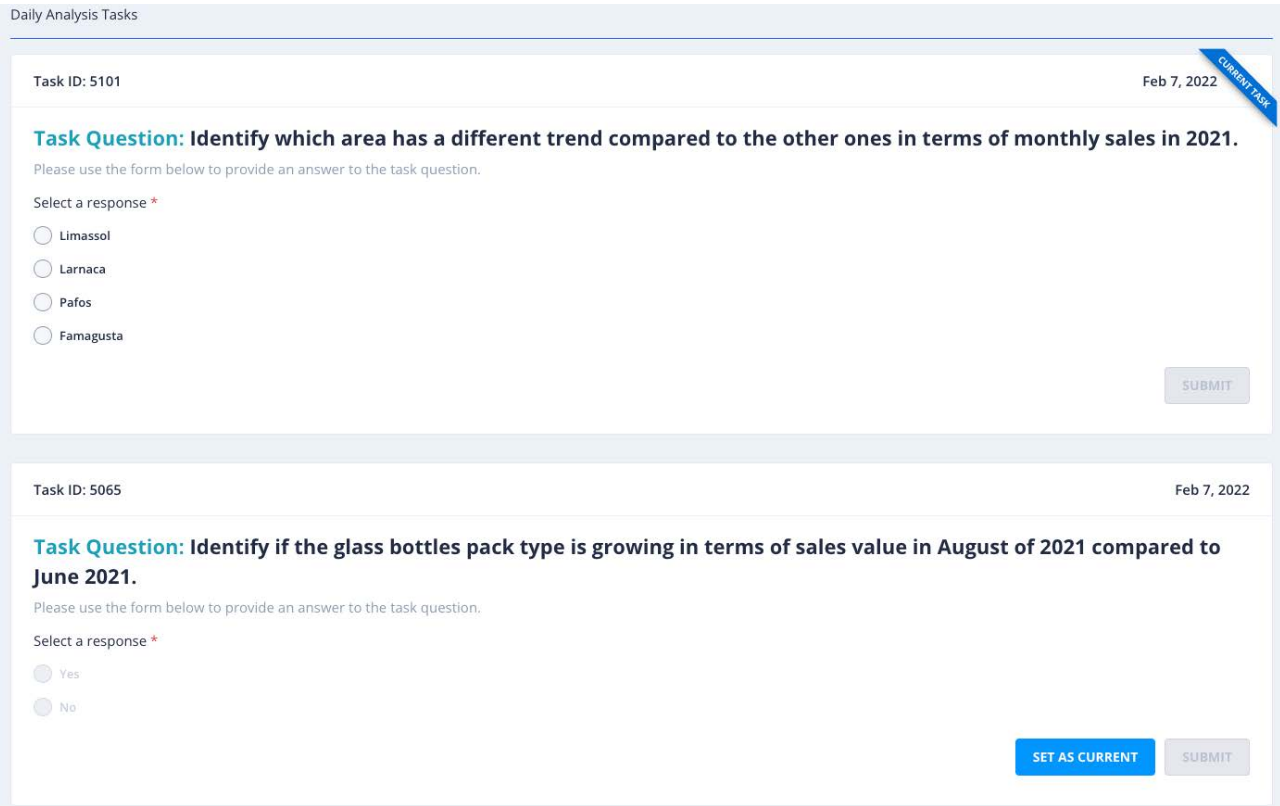



*Figure 2 - Participant's Interface for Initiating or Responding to an Analysis Task*

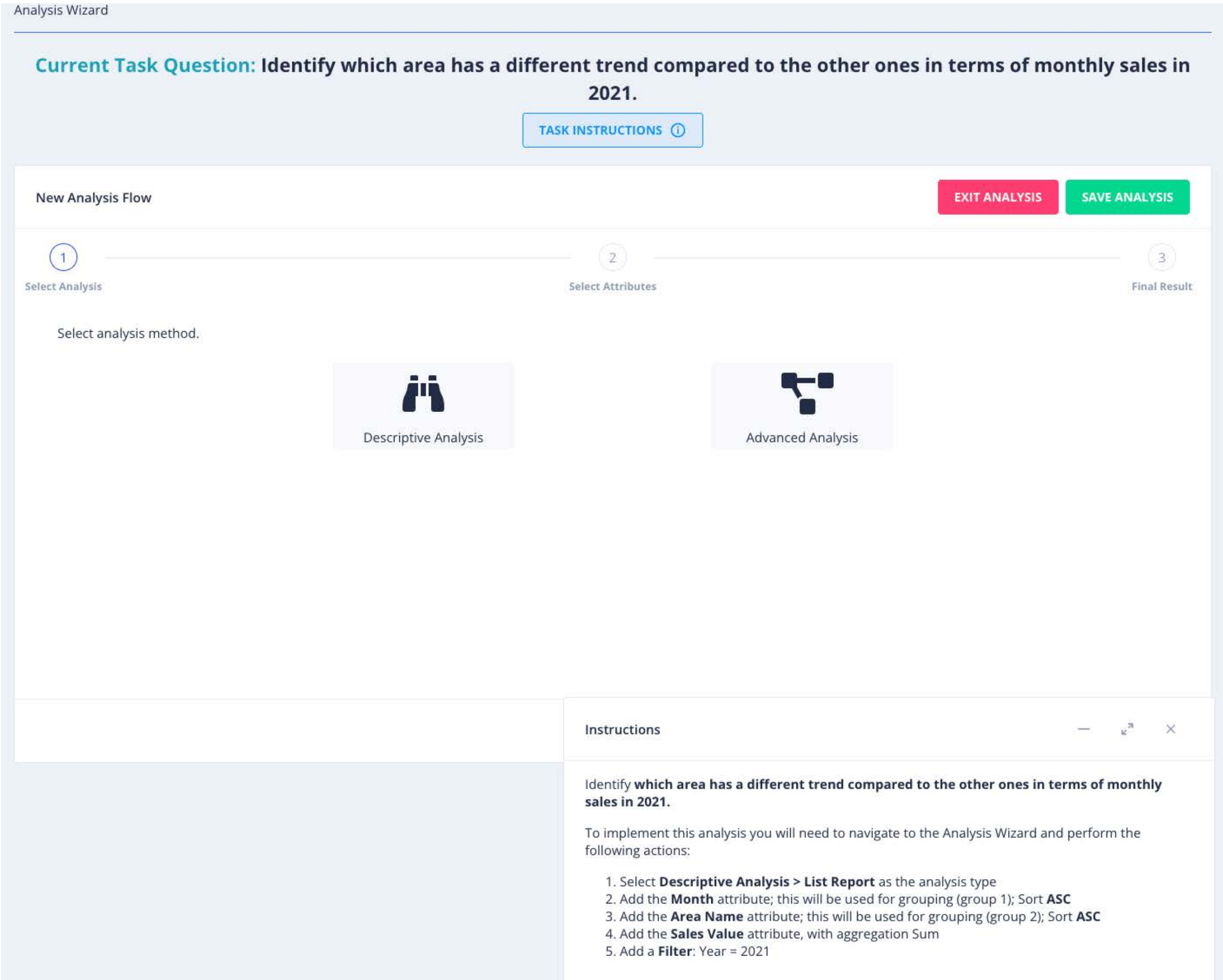



*Figure 3 - Analysis Task Modal of Instructions*

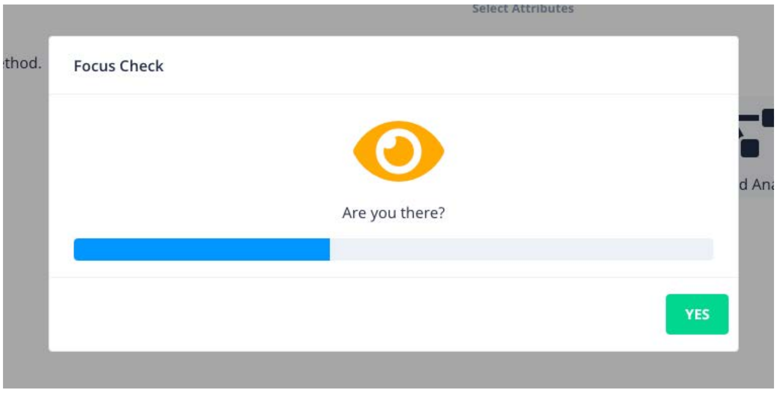



*Figure 4 - Mechanism that Ensures Focused Participation*

### 3.1.3 PROCEDURE USED

Due to COVID-19 all operations regarding the pilot study had to be performed remotely. After communicating with the collaborator organizations, the team was able to secure the recruitment of 67 data analyst for the study, some of who were new to the project (i.e., did not participate in previous user studies). In what follows we further discuss the study procedure .

**Participant Training:** All participants were invited to a remote MS Teams meeting where the project team introduced the project. The onboarding phase was done to make sure that new participants were up to speed with the project's direction and goals. Moreover, the study's use case was presented, informing participants of their role as brand managers and their assigned task which was to utilise the IDEALVis platform to perform a set of tasks regarding their brand called IdealCola. Moving on, during the training session the project team introduced all participants to the platform and demonstrated how an analysis task can be addressed. After the system/study demonstration, the dates of the study were announced to participants. Towards the end of the training session the participants were given time to ask questions regarding the study. Once the training session was concluded, a recorded version of the presentation and system demonstration videos was sent to all participants, so they could revisit the training material covered. Finally, the project team sent an email to each individual participant with the URL to the platform along with their personal user credentials for logging in to the platform. This allowed the participants to use the platform during their own time and at their own pace to practice the set of analysis tasks that were demonstrated during training. This was done to ensure that all participants familiarised themselves with the overall interface and analysis tools of the platform prior to the actual study. It should be noted that when participants first logged in to the system, they received a welcome message along with a presentation demonstrating all the features of the IDEALVis system, including both study features and also GDPR related features, such as enabling a participant to request the deletion of their information in case they wished to do so. This presentation of features was also available later during the actual study, if the participant wished to revisit it.

**New Participant Onboarding:** For the new participants that had never participated in any of the IDEALVis user studies, an onboarding procedure was set in place. Specifically, all new participants were contacted via email with specific instructions on how to complete their IDEALVis user profile. The completion of the user profile was a requirement for the pilot study, as it formed the basis for acquiring the user model of each participant that was necessary for adapting data visualizations.

**Study Part A:** Once participants completed the set of demonstration tasks and the onboarding process (i.e., assuming a new participant), they were able to access the first set of 19 analysis tasks (Part A of the study). Each participant had to utilise the IDEALVis system to address these 19 tasks, that were the control tasks i.e., the resulting data visualization type was predefined for each task and not adapted/personalized. Participants were given a total of 6 days to complete all 19 analysis tasks from Part A of the study. Once participants were done with Part A, they were asked to complete a questionnaire (more on this in Section 4).

**Study Part B:** Once participants completed the analysis tasks and the provided questionnaire from Part A of the study, they were able to access the second set of 19 analysis tasks (Part B of the study). Similar to Part A, the participants were required to utilise the IDEALVis system to address these 19 tasks, that were the adaptive tasks i.e., the resulting data visualization type for each task was adapted/personalized according to the participant's user model. Participants were given 15 days to complete all 19 analysis tasks from Part B. More time was given for Part B to accommodate for participants who delayed with the completion of Part A. Once participants were done with Part B, they were asked to complete another questionnaire (more on this in Section 4).

**Addressing a Task:** The process of addressing a task in both study parts (i.e., Part A and Part B) was the same and it was comprised of 6 steps.

- Step 1: the participants had to follow a list of all available analysis tasks for Part A or Part B, respectively. The tasks were presented in a random order, so that no two participants follow the same order of addressing the set of analysis tasks.
- Step 2: participants had to select one of the tasks as being their current task (Figure 2).
- Step 3: participants had to navigate to the Analysis Wizard interface to begin the exploration process (the current analysis task narrative was always on the top of the Analysis Wizard).
- Step 4: participants had to complete the three steps of the Analysis Wizard (select analysis step, select attributes step, and view result step) according to the instructions provided in the instructions modal (Figure 3).
- Step 5: following the final step of the Analysis Wizard (Step 4), participants had to review and understand the resulting visualisation (adapted or not depending on the part of the study) and accordingly form a response to the task question (decide their response which they provide it in step 6). It should be noted that this is the step where the performance of the participants in terms of view time in milliseconds is recorded.
- Step 6: participants had to navigate back to the list of analysis tasks and provide their answer to the current task. This is the step where the accuracy of the participant is recorded.

For the analysis process participants were informed that once they start the exploration process (i.e., analysis for a specific task), they cannot stop until the specific task is addressed i.e., a response to the task is provided.

## 3.2 *Pilot Study Analysis Results*

For the pilot study collected performance and accuracy responses from 45 participants for all 38 analysis tasks i.e., 19 control tasks with non-adapted data visualizations and 19 tasks with adapted/personalised data visualizations. While our initial sample of participants was 67, we omitted some of the participants from our sample, since they were not able to complete the pilot study in the timeframe provided. This section reports the IDEALVis platform’s evaluation results with regards to enabling participants achieve a more efficient and effective data exploration of business datasets through the use of adapted/personalised data visualizations).

### 3.2.1 PERFORMANCE FINDINGS

In this section we explore the impact of data visualization adaption on the participants' task performance i.e., time taken for the participants' (in terms of milliseconds) to address an analysis task. For analysing performance, we only used response records of task pairs where the participant responded accurately to both analysis tasks i.e., the participant's response was valid for both related tasks across the two study conditions (i.e., task with adaptation disabled and task with adaptation enabled). During our analysis we assessed that for task pair T13 (Table 3) all participants took an extreme amount of time to complete the personalised variant of the task. After consulting with the participants about that matter we found out that most participants were struggling to find the answer to this task / or they were not sure how to exactly approach it. Some participants even reported having to use a calculator to find the correct answer. A similar pattern was detected for the performance records of task pair T12. Accordingly, we decided to exclude these task pairs from the performance analysis presented in this section. The analysis across the two study conditions revealed that adaptation had a positive effect on participants' performance enabling them to achieve an average decrease of 8.1 ± 6.9 seconds with regards to task completion time. Moreover, with adaptation enabled, performance improved for an average of 9 ± 2 tasks per participant, while the number of tasks improved in terms of performance at the unique participant level was at maximum 15 tasks and at minimum 5 tasks. Additionally, with adaptation enabled, performance worsen for an average of 2 ± 1 tasks per participant, while the number of tasks worsen in terms of performance at the unique participant level was at maximum 5 tasks and at minimum 0 tasks.

Analysis on the impact of adaptation with regards to performance across different task types shows that adaptation had a positive effect on participants' performance enabling them to achieve (i) a statistically significant average decrease of 7.8 seconds for Retrieve Value tasks ($p < .01$), (ii) a statistically significant average decrease of 25.9 seconds for Correlation tasks ($p = .01$), (iii) a statistically significant average decrease of 8.2 seconds on Simple Comparison tasks ($p < .01$) and (iv) a non-statistically significant average decrease of 10.6 seconds on Compute Derived Value tasks ($p = 0.24$). Since Simple Comparison tasks was the larger group of analysis tasks (10 task pairs), we decided to further explore this group of tasks by independently analysing Simple Comparison tasks which used time series data. Results show that with adaptation enabled participants achieved (i) a statistically significant average decrease of 9.9 seconds on Simple Comparison tasks which used time series data ($p < .01$) and (ii) a statistically significant average decrease of 4.5 seconds on the remaining Simple Comparison tasks ($p < .01$). Moving on, with adaptation enabled, performance improved for an average of 84 ± 82 task responses across all analysis task types, while the number of task responses improved in terms of performance at the unique analysis task type level was at maximum 199 responses for Simple Comparison tasks which used time series data, and at minimum 5 responses for Compute Derived Value tasks.

Finally, we report that with adaptation enabled, performance worsen for an average of 22 ± 21 task responses across all analysis task types, while the number of task responses that worsen in terms of performance at the unique analysis task type level was at maximum 52 responses for Simple Comparison tasks which used time series data and at minimum 1 response for Correlation and Compute Derived Value tasks. Unfortunately, the sample of our tasks was limited to a single Find Anomaly pair of tasks, for which most participants were only able accurately respond to the personalised variant of the task, leaving only a small sample that was considered very small to contribute any valid results to the analysis.

#### 3.2.2 ACCURACY FINDINGS

In this section we explore the impact of data visualization adaption on the participants' accuracy i.e., the participants' ability to address a specific analysis task correctly. For each of the study task conditions a participant was able to achieve a maximum score of 19 since each condition had a set of 19 tasks. Analysing the accuracy scores of each participant reveals that 62% of participants were more accurate when addressing analysis tasks with adapted/personalised data visualizations.

Moreover, 18% of participants were not affected in terms of accuracy across the two study conditions, while the remaining 20% of participants were negatively impacted by adaptation in terms of accuracy. In contrast to analysis tasks with no data visualization adaptation, participants were able to address on average an additional 8% of analysis tasks correctly when working with tasks delivering data visualization adaptation. Analysis of accuracy scores across task types for both conditions revealed that participants were generally much more accurate in addressing tasks when adaptation was enabled for Simple Comparison, Compute Derived Value and Find Anomaly tasks. Specifically, participants were more accurate by 6.6% for Simple Comparison tasks, 34.2% for Computer Derived Value tasks and 90% for Find Anomaly tasks. In contrast, for Correlation and Retrieve Value task types we were not able to see a significant impact in terms of accuracy when participants were using adapted/personalised data visualizations for addressing the analysis tasks.

# 4 Adaptation's Impact on User Experience and System Usability Factors

In this section we explore the analysis results regarding the evaluation of the platform's user experience and usability factors. During the pilot study (as mentioned is Section 3) the participants of the study were exposed to two sets of analysis tasks they had to explore. The first set of analysis tasks required participants to address each task by exploring a dataset using data visualizations that were not adapted to the participant's characteristics (i.e., user model). We refer to this as pilot Part A. Moreover, analysis tasks in the second set had a similar nature to those of the first set i.e., in terms of task complexity and task type. Instead for this second set of analysis tasks when the participant was exploring the dataset to address a specific task, the requested data visualizations were adapted automatically by the system according to the unique participant's user model. We refer to this as pilot Part B. In the next sections, we demonstrate the procedure in which the user experience and usability factors were collected during the pilot study and then we summarise the analysis results of those factors, focusing on how the factors' scores were impacted by the adaptation offered by the system.

## 4.1 Procedure Used

For being able to understand the impact of adaptation on user experience and usability factors, we had to capture the participant's views regarding these factors at two distinct phases. Specifically, once a participant had successfully responded to all analysis tasks of the first set of tasks i.e., pilot Part A, a link to a questionnaire measuring the systems user experience and usability factors was sent to them by the team. Once a participant had successfully completed this questionnaire, the second set of analysis tasks were added for them on the IDEALVis platform so they could move to pilot Part B. Moving on, when a participant had successfully completed Part B of the pilot's analysis tasks, they were invited to participate in an identical questionnaire measuring the same factors (i.e., user experience and system usability).

### 4.1.1 MATERIALS USED

For being able to capture the user experience and system usability factors for the IDEALVis platform we utilised two accredited system evaluation questionnaires which we combined into a web-based questionnaire that was forwarded to our participants at two distinct phases as mentioned above. Specifically for measuring the participants' user experience with the system we used the User Experience Questionnaire Sort Version (UEQ-S) (UEQ, 2022). According to the questionnaire's authors, this questionnaire's scales "*cover a comprehensive impression of user experience. Both classical usability aspects (efficiency, perspicuity, dependability) and user experience aspects (originality, stimulation) are measured*". Moreover, for measuring the system's usability we used the System Usability Scale (SUS) questionnaire. This 10-scale questionnaire provides a reliable tool for measuring the usability of a system. We chose this tool as its deemed appropriate for our purpose since it has become an industry standard, with references in over 1300 articles and publications (Usability.gov, 2022). The two questionnaire scales can be seen in Figure 5 and Figure 6. Below we also provide links to the two web-based system evaluation questionnaires forwarded to participants during the pilot study.

**Evaluation Questionnaire completed after Part A: Link**

**Evaluation Questionnaire completed after Part B: Link**

English version

| | | |
|---|---|---|
| obstructive | o o o o o o o | supportive |
| complicated | o o o o o o o | easy |
| inefficient | o o o o o o o | efficient |
| confusing | o o o o o o o | clear |
| boring | o o o o o o o | exciting |
| not interesting | o o o o o o o | interesting |
| conventional | o o o o o o o | inventive |
| usual | o o o o o o o | leading edge |

*Figure 5 - Short Version of the User Experience Questionnaire (UEQ)*

| | The System Usability Scale<br>Standard Version | Strongly Disagree<br>1 | 2 | 3 | 4 | Strongly Agree<br>5 |
|---|---|---|---|---|---|---|
| 1 | I think that I would like to use this system frequently. | O | O | O | O | O |
| 2 | I found the system unnecessarily complex. | O | O | O | O | O |
| 3 | I thought the system was easy to use. | O | O | O | O | O |
| 4 | I think that I would need the support of a technical person to be able to use this system. | O | O | O | O | O |
| 5 | I found the various functions in this system were well integrated. | O | O | O | O | O |
| 6 | I thought there was too much inconsistency in this system. | O | O | O | O | O |
| 7 | I would imagine that most people would learn to use this system very quickly. | O | O | O | O | O |
| 8 | I found the system very awkward to use. | O | O | O | O | O |
| 9 | I felt very confident using the system. | O | O | O | O | O |
| 10 | I needed to learn a lot of things before I could get going with this system. | O | O | O | O | O |

The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

*Figure 6 - System Usability Scale Questionnaire*

## *4.2 Analysis Results for User Experience*

For the User Experience Questionnaire, we were able to collect responses from a total of 35 participants, both after pilot Part A and after pilot Part B. Moreover, the analysis of these responses was handled by an automated process offered by the questionnaire's authors. The UEQ questionnaire contains 8 scales that need to be answered by each participant, with each scale taking a value from 1 to 7. Moreover, the first 4 scales in this questionnaire are used to measure the pragmatic quality metric while the rest 4 scales measure the hedonic quality metric.

**Pragmatic Quality:** This metric focuses on the task-oriented nature of an experience. For example, this considers the task's efficiency and ease of use etc.

**Hedonic Quality:** This metric focuses more on the fun, appeal and more generally on the originality aspects of the experience offered by a system.

Using the responses of all participants we calculate the Cronbach's alpha (or coefficient alpha) for each set of scales belonging to each metric i.e., pragmatic quality and hedonic quality. It is expected that scales that belong to the same group should show in general a high correlation and therefore we use the Cronbach's alpha (Cronbach, 1951) which is a measure for the consistence of a scale. In general, an alpha value of more than 0.7 is usually considered acceptable. Performing this statistical calculation also helps us understand that the different scales of the questionnaire were interpreted as intended by the participants. In our results we do not mention Cronbach's alpha results since those were acceptable (i.e., alpha > 0.7) for both pragmatic and hedonic quality scales for data collected from Part A and Part B of the pilot. In the next sections, we provide the user experience results for both pilot phases and we further explore the impact of adaptation with regards to the participants' user experience when using the IDEALVis platform to perform data analysis tasks.

### 4.2.1 USER EXPERIENCE RESULTS (PART A – ADAPTATION DISABLED)

| Item | Mean | Variance | Std. Dev. | No. | Negative | Positive | Scale |
|---|---|---|---|---|---|---|---|
| 1 | 1.6 | 1.3 | 1.1 | 35 | obstructive | supportive | Pragmatic Quality |
| 2 | 1.0 | 2.1 | 1.4 | 35 | complicated | easy | Pragmatic Quality |
| 3 | 1.5 | 1.2 | 1.1 | 35 | inefficient | efficient | Pragmatic Quality |
| 4 | 1.4 | 1.6 | 1.3 | 35 | confusing | clear | Pragmatic Quality |
| 5 | 0.7 | 2.9 | 1.7 | 35 | boring | exciting | Hedonic Quality |
| 6 | 0.9 | 3.0 | 1.7 | 35 | not interesting | interesting | Hedonic Quality |
| 7 | 1.0 | 1.2 | 1.1 | 35 | conventional | inventive | Hedonic Quality |
| 8 | 0.8 | 1.0 | 1.0 | 35 | usual | leading edge | Hedonic Quality |

*Figure 7 - UEQ Scores per Scale (Part A - Adaptation Disabled)*

| Short UEQ Scales | |
|---|---|
| Pragmatic Quality | 1.357 |
| Hedonic Quality | 0.864 |
| Overall | 1.111 |

*Figure 8 -UEQ Final Platform Results (Part A - Adaptation Disabled)*

### 4.2.2 USER EXPERIENCE RESULTS (PART B – ADAPTATION ENABLED)

| Item | Mean | Variance | Std. Dev. | No. | Negative | Positive | Scale |
|---|---|---|---|---|---|---|---|
| 1 | 1.6 | 0.9 | 0.9 | 35 | obstructive | supportive | Pragmatic Quality |
| 2 | 1.4 | 1.3 | 1.1 | 35 | complicated | easy | Pragmatic Quality |
| 3 | 1.3 | 1.5 | 1.2 | 35 | inefficient | efficient | Pragmatic Quality |
| 4 | 1.5 | 1.8 | 1.3 | 35 | confusing | clear | Pragmatic Quality |
| 5 | 0.8 | 1.9 | 1.4 | 35 | boring | exciting | Hedonic Quality |
| 6 | 0.9 | 2.0 | 1.4 | 35 | not interesting | interesting | Hedonic Quality |
| 7 | 1.2 | 0.8 | 0.9 | 35 | conventional | inventive | Hedonic Quality |
| 8 | 1.0 | 0.7 | 0.8 | 35 | usual | leading edge | Hedonic Quality |

*Figure 9 - UEQ Scores per Scale (Part B - Adaptation Enabled)*

| Short UEQ Scales | |
|---|---|
| Pragmatic Quality | 1.450 |
| Hedonic Quality | 0.979 |
| Overall | 1.214 |

*Figure 10 - UEQ Final Platform Results (Part B - Adaptation Enabled)*

### 4.2.3 INTERPRETATION OF UEQ RESULTS

According to the authors of the User Experience Questionnaire "*Values between -0.8 and 0.8 represent a neural evaluation of the corresponding scale, values > 0.8 represent a positive evaluation and values < -0.8 represent a negative evaluation*". Taking into consideration the results of both Part A and Part B we can see that most scales for user experience had a value of above 0.8 therefore we conclude that overall, the user experience evaluation was generally positive with or without adaptation/personalization (Figure 7 and Figure 9). The only scale that was rated below 0.8 was that of the boring/exciting scale which is one of the hedonic quality scales. Moreover, the occurrence of this neutral evaluation was in the responses provided for the non-adapted visualizations in pilot Part A. Interestingly enough, this specific scale of boring/exciting was increased by 0.1 when adaptation was enabled finally reaching a score of 0.8.

Moving on, we discuss the results in terms of the impact of data visualization adaptation/personalization on user experience. As expected, most scales of both pragmatic and hedonic quality were evaluated higher by participants after they engaged with the adapted/personalised data visualizations. The only scale that was slightly decreased (just by 0.2) after adaptation was enabled is that of inefficient/efficient which is one of the pragmatic quality scales. We do not consider this decrease significant as it is only related to a specific isolated scale which was not sufficient to affect the overall score of pragmatic quality. Instead as the results show (Figure 8 and Figure 10) pragmatic and hedonic qualities were both significantly increased when the adaptation/personalization condition was enabled. Specifically, when participants received personalised data visualizations the reported pragmatic quality was increased by almost 0.1, hedonic quality was increased by 0.11, while the overall evaluation score increased by 0.1. The user experience evaluation results captured for both the adapted and non-adapted parts of the pilot study, are visualised in Figure 11 to further demonstrate the impact of adaptation/personalization on data visualizations with regards to user experience.

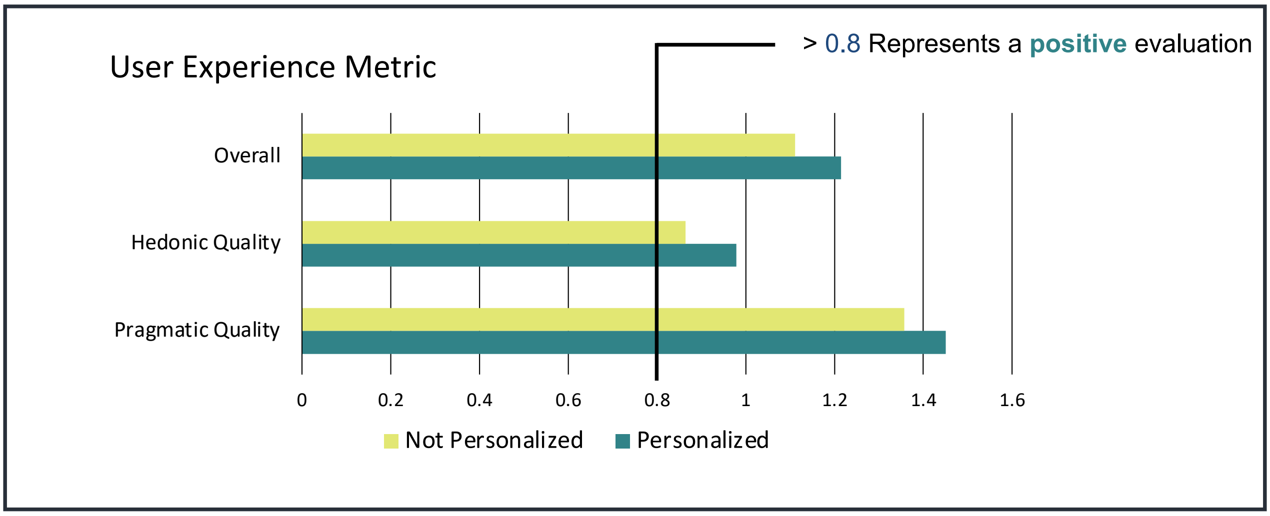



*Figure 11 - Increase in User Experience Metrics When Visualizations are Adapted/Personalised*

### *4.3 Analysis Results for System Usability*

For the System Usability Scale Questionnaire, we were able to collect responses from a total of 35 participants, both after pilot Part A and after pilot Part B. Moreover, the analysis of these responses was handled by an automated process offered by the questionnaire's authors. In the next sections, we provide the system usability results for both pilot phases, and we further explore the impact of adaptation with regards to system usability when using the IDEALVis platform to perform data analysis tasks.

### 4.3.1 SYSTEM USABILITY SCORE RESULTS (PART A – ADAPTATION DISABLED)

The average System Usability Score for the platform when participants interacted with the non-adapted data visualizations was 66.2 with a standard deviation of 12.9. As seen in Figure 12 this score denoted that the usability of the platform is considered marginally acceptable.

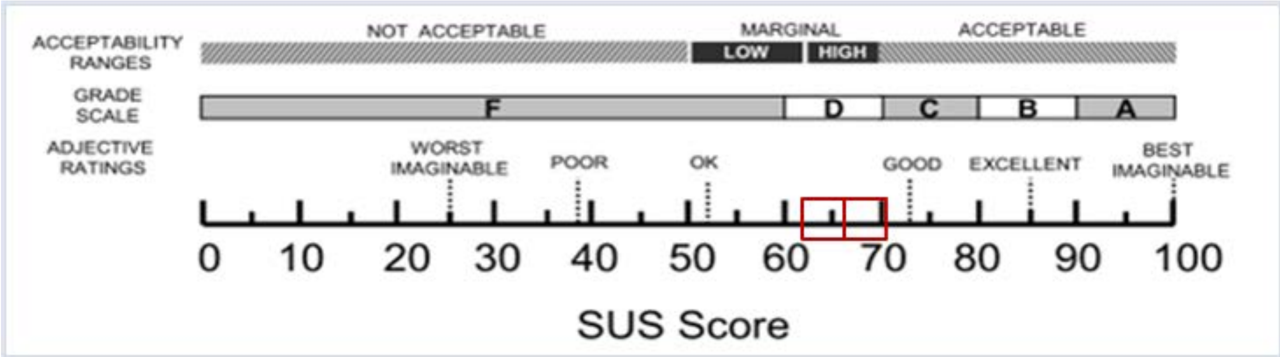



*Figure 12 - System Usability Score (Part A – Adaptation Disabled)*

### 4.3.2 SYSTEM USABILITY SCORE RESULTS (PART B – ADAPTATION ENABLED)

The average System Usability Score for the platform when participants interacted with the adapted data visualizations was 67.4 with a standard deviation of 11. Again, this score denotes that the usability of the platform is considered marginally acceptable.

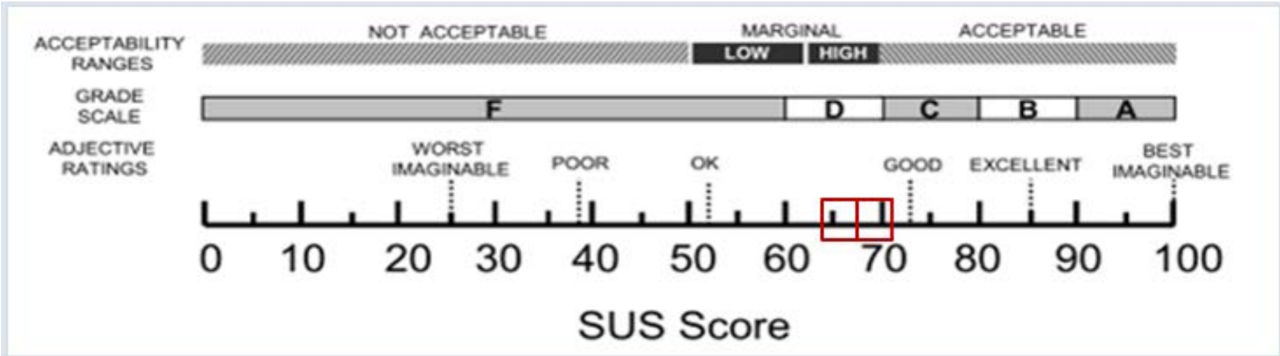



*Figure 13 - System Usability Score (Part B – Adaptation Enabled)*

### 4.3.3 INTERPRETATION OF SUS RESULTS

While the usability scores achieved by the platform across the two conditions (i.e., data visualization adaptation enabled / disabled) is marginal, the usability score increased by 1.2 after participants interreacted with the tasks which implemented adapted/personalised data visualizations. Additionally, another important aspect is that the standard deviation of the system usability score after participants were exposed to adapted data visualizations is lower than the standard deviation of the system usability score representing Part A where there was no adaptation. A smaller standard deviation means that the average usability scores elicited from the participants' responses are closer to the mean and thus we are more confident for the higher system usability score achieved after adaptation was enabled. Finally, a usability score of above 68 is considered above average (Sauro, 2022). The overall evaluation of IDEALVis platform revealed that enabling the adaptation conditions when end-users interact with the given tasks facilitated an increase of their perceived usability reaching to a marginal value closer to that of the average score. The latter could be considered acceptable for the first release of the IDEALVis platform considering the peculiarities and complexity of the business domain.

# 5 Discussion

Our evaluation user study shows that the IDEALVis platform was able to positively affect the participants' perceived user experience and perceived system usability scores, but most importantly was able to improve the participants' performance and accuracy across a variety of data analysis tasks.

While our work committed to the improvement of the overall efficiency and effectiveness of the business data analyst when addressing data analysis tasks, there are some limitations that we would like to address in the future. The sample of analysis tasks used during evaluation was not balanced in terms of task type since more focus was given on simpler comparison tasks. Moreover, this work did not report results with regards to which adaptations/interventions were the direct enablers for the participants' improvement in terms of accuracy and performance.

Some questions rising from this work that we plan in addressing as part of future endeavours includes: (i) How could our approach offer a transparent explanation to the business analyst as with regards to why the best-fit data visualization was selected? (ii) How can we more effectively process the resulting user's interaction with the adapted output and further gain insight on which adaptation/intervention was the most helpful for that type of user? and (iii) How does our adaptation perform with unexplored data visualizations and analysis task types? Our goal is to attempt to address these questions in several ways. We plan to extend our sample of users by applying this work to more industry domains and gathering more data visualization interaction data that can yield more diverse adaptation rules. In this way will gain a deeper understanding of the impact of other human factors on data visualizations and their explorations to improve the IDEALVis adaptation engine.

# 6 Conclusions

This deliverable presented the overall evaluation performed for the IDEALVis platform. The deliverable carefully laid out the goals and the excepted outcomes of IDEALVis and based on those it defined the metrics with which the evaluation of this platform was carried out. Moreover, throughout the deliverable we have seen the setup, design, and procedure of the pilot study as well as the different type of materials / data collection procedures used during the pilot study, required for capturing all the appropriate metrics essential for the platform's evaluation. Finally, the analysis results of the different metrics captured during the pilot study were presented with emphasis given on how data visualization adaptation offered by IDEALVis was able to positively influence those evaluation metrics in helping the participants achieve a more effective and efficient data analysis of business data.

# References

Cronbach, L. J., 1951. Coefficient alpha and the internal structure of tests. *Psychometrika,* 16(3), pp. 297-334.

Sauro, J., 2022. *Measuring U.* [Online] Available at: https://measuringu.com/sus/ [Accessed 20 03 2022].

UEQ, 2022. *User Experience Questionnaire.* [Online] Available at: https://www.ueq-online.org [Accessed 05 03 2022].

Usability.gov, 2022. *Usability.gov.* [Online] Available at: https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html [Accessed 05 03 2022].