

ABC TRAINING CENTER

A minor project reports

Submitted in partial fulfillment of the requirements for the award of degree of

**BACHELOR OF TECHNOLOGY
(COMPUTER SCIENCE AND ENGINEERING)**

Submitted to

**LOVELY PROFESSIONAL UNIVERSITY
PHAGWARA, PUNJAB**



From 10/03/23 to 09/04/23

SUBMITTED BY

Name of student: SHIVAM VERMA

Registration Number: 12017149

Signature of the Student:

Student Declaration

To whom so ever it may concern

I, Shivam Verma, 12017149, hereby declare that the work done by me on “R Programming” on topic of “ABC Training Center” from 10th March 2023 to 09th April 2023, is a record of original work for the partial fulfillment of the requirements for the award of the degree, Bachelor of Technology.

SHIVAM VERMA (12017149)



Signature of the student

Dated: 9th April 2023.

Acknowledgement

Primarily I would like to thank God for being able to learn a new technology. Then I would like to express my special thanks of gratitude to the teacher and instructor of the course R programming who provide me the golden opportunity to learn a new technology from home.

I would like to also thank my own college, Lovely Professional University for offering such a course which not only improve my programming skill but also taught me other new technology.

Then I would like to thank my parents and friends who have helped me with their valuable suggestions and guidance in choosing this course.

Last but not least I would like to thank my all classmates who have helped me a lot.

Date: 09/04/2023

Shivam Verma

Reg no: 12017149

Table of Contents

S. No.	Title	Page
1	Cover Page	1
2	Declaration of the student	2
3	Acknowledgement	3
4	Table of Contents	4
5	Introduction	5
6	Scope Of The Analysis	6
7	Drawback Of Existing Technology	7
8	Source Of dataset	8
9	Introduction of dataset	9
10	General Description	11
11	Specific Requirement	12
12	Analysis Result	15
13	Visualization	21
14	List Of Analysis with Result	22
15	Future Scope	23
16	References	25
17	Bibliography	26

Introduction

R programming is a widely used open-source statistical programming language and environment for data analysis and visualization. It provides a rich set of tools for data manipulation, analysis, and visualization, making it a popular choice among data scientists, statisticians, and analysts. R programming is known for its flexibility, ease of use, and powerful capabilities for data analysis, making it an essential skill for professionals in various fields.

In recent years, the demand for R programming skills has grown exponentially due to the increasing need for data-driven decision making in industries such as finance, healthcare, marketing, and technology. As a result, many educational institutions, including ABC Training Center, have developed comprehensive training programs to introduce learners to the fundamentals of R programming and equip them with the necessary skills to use R for data analysis and visualization.

The R programming course at ABC Training Center is designed to provide learners with a solid foundation in R programming, covering topics such as basic syntax, data types, and data structures in R, data manipulation techniques, data analysis methods, and real-world data analysis projects. The course follows a learner-centered instructional approach that combines theory with hands-on learning, allowing learners to practice their skills through interactive discussions, demonstrations, and practical exercises. The course is taught by expert instructors who provide guidance and support to learners throughout the program.

This module provides an overview of R programming language, its syntax, and basic data types and structures in R.

Data manipulation in R: This module covers various techniques for manipulating data in R, including data importing, data cleaning, data transformation, and sub setting.

Data analysis with R: This module focuses on data analysis techniques using R, such as data aggregation, statistical analysis, and data visualization using R's built-in functions and packages.

Advanced topics in R programming: This module covers more advanced topics in R programming, such as working with data frames, writing custom functions, and using conditional statements and loops.

Real-world data analysis projects: This module provides hands-on experience in applying R programming skills to real-world data analysis projects, allowing learners to practice their skills and gain practical experience.

Scope Of The Analysis

The scope of analysis in the R programming course at ABC Training Center is broad and covers various aspects of data analysis and visualization. The course aims to provide learners with a solid foundation in R programming, equipping them with the skills to manipulate, analyze, and visualize data using R's built-in functions and packages.

The scope of analysis in the R programming course may include:

- **Basic Syntax and Data Types:** Learners will learn the fundamental syntax and data types in R, including variables, data frames, and vectors. They will also gain an understanding of R's data structures and how to manipulate data using R's built-in functions.
- **Data Manipulation:** Learners will learn techniques for data manipulation, including data cleaning, data aggregation, merging data frames, and reshaping data. They will also learn how to handle missing data and deal with data inconsistencies.
- **Data Analysis:** Learners will be introduced to various data analysis techniques, such as descriptive statistics, exploratory data analysis (EDA), data visualization, and basic statistical tests using R's statistical functions and packages.
- **Data Visualization:** Learners will learn how to create visualizations using R's graphical libraries, such as ggplot2, to effectively communicate insights from data. They will learn to create various types of plots, such as bar charts, line charts, scatter plots, and heatmaps.
- **Real-world Data Analysis Projects:** Learners will have the opportunity to apply their R programming skills to real-world data analysis projects. They will learn how to load and clean data, perform data analysis tasks, create visualizations, and interpret results to derive insights and make data-driven decisions.
- **The scope of analysis in the R programming course at ABC Training Center is designed to provide learners with a comprehensive understanding of R programming for data analysis and visualization. The course covers essential topics and techniques that are commonly used in the field of data analysis and prepares learners to apply their skills to real-world scenarios.**

Drawback Of Existing Technology

Like any other technology, R programming also has some limitations or drawbacks. Here are some potential limitations of R programming:

- **Steeper Learning Curve:** R programming has a steeper learning curve compared to some other programming languages, especially for learners who are new to programming. R has its own unique syntax and data structures, which may take some time to grasp for beginners without prior programming experience.
- **Memory Management:** R is an interpreted language, which means that it may not be as efficient in terms of memory management as compiled languages like C or Java. This can lead to performance issues when working with large datasets or performing computationally intensive tasks.
- **Limited Scalability:** While R is powerful for data analysis and visualization, it may not be the best choice for large-scale or production-level applications. R is primarily designed for interactive data analysis and may not be as efficient for handling very large datasets or high-performance computing tasks.
- **Limited Support for Multi-threading:** R's core functions are not inherently designed for multi-threading, which can limit its performance in multi-core processor environments. This can result in slower performance when dealing with computationally intensive tasks that could otherwise benefit from parallel processing.
- **Package Fragmentation:** R has a large ecosystem of packages contributed by the community, which can sometimes result in fragmentation and inconsistency in terms of documentation, functionality, and compatibility. This can make it challenging for users to find the right packages for their specific needs and navigate the vast package ecosystem.
- **Relatively Slower Execution Time:** Due to its interpreted nature, R may not be as fast in terms of execution time compared to compiled languages like C or Java. This can impact the performance of computationally intensive tasks or large-scale data processing.
- **Limited Support for Object-Oriented Programming (OOP):** R is primarily a functional programming language and does not have robust support for object-oriented programming (OOP) compared to some other programming languages. This can limit the flexibility and reusability of code in certain scenarios.

It's worth noting that despite these limitations, R programming has a strong community support, a vast ecosystem of packages, and a wide range of applications in data analysis, statistics, and visualization. Many of these limitations can be mitigated through optimization techniques, parallel computing, and using optimized packages. Additionally, R continues to evolve with regular updates and improvements, addressing some of these limitations over time.

Source Of Dataset

The source of datasets used in R programming can vary depending on the specific analysis or project requirements. Some common sources of datasets for R programming include:

1. **Publicly Available Datasets:** There are many publicly available datasets that can be used for data analysis in R. These datasets can be obtained from various sources, such as government agencies, research institutions, open data portals, and data repositories. Examples of popular public datasets include the Iris dataset, Titanic dataset, and the Boston Housing dataset.
2. **Real-world Datasets:** Datasets obtained from real-world sources, such as industry or business datasets, can be used for data analysis in R. These datasets may be proprietary or obtained through collaborations with organizations or companies, and they can provide valuable insights and practical applications of R programming in real-world scenarios.
3. **Custom Generated Datasets:** Datasets can also be generated or simulated within R programming itself using functions or packages that provide data generation capabilities. This allows for the creation of custom datasets with specific characteristics or distributions for analysis or modeling purposes.
4. **Kaggle and Other Data Science Platforms:** Kaggle, a popular online platform for data science competitions, provides a vast repository of datasets that can be used for data analysis in R. Other data science platforms and communities may also offer datasets for analysis or learning purposes.
5. **APIs and Web Scraping:** R has capabilities to interact with APIs (Application Programming Interfaces) to access data from various online sources, such as social media platforms, financial data providers, weather APIs, etc. Web scraping is another technique that can be used to extract data from websites and convert it into a format that can be used for data analysis in R.

It's important to note that when using datasets in R programming, it's essential to comply with any data usage policies, permissions, and ethical considerations associated with the data source. Proper data handling and privacy practices should be followed to ensure the integrity and legality of data usage.

Introduction

Welcome to the training data management Excel file! This spreadsheet is designed to help you efficiently organize and analyze training data for your organization. It includes columns for essential information such as Training ID, Name, Date, Package Price, Participants, Revenue, Training Location, and Training Type and having 5000 records.

The Training ID column serves as a unique identifier for each training session, making it easy to track and manage different training events. The Name column allows you to input the names of the training sessions or programs for easy reference. The Date column enables you to record the date on which each training session took place.

The Package Price column provides space for recording the cost of each training package, allowing you to easily calculate the total revenue generated from training sessions. The Participants column lets you input the number of participants who attended each training session, which can be used to analyze attendance trends and evaluate the success of different training programs.

The Revenue column automatically calculates the total revenue generated based on the package price and number of participants, providing you with accurate financial information. The Training Location column allows you to input the location where each training session took place, making it easy to track and manage training venues.

Finally, the Training Type column provides space for recording the type of training session, whether it's a workshop, seminar, webinar, or any other type of training. This information can be used to analyze the effectiveness of different training formats and make informed decisions for future training programs.

Overall, this Excel file is a comprehensive tool for managing and analyzing training data, providing you with valuable insights to optimize your training programs and improve the overall effectiveness of your organization's training initiatives.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	
1	Training ID	Date	Name	Package Price	Participants	Revenue	Training Location	Training_Type													
2	AT-005	01-Jan-18	2D Animation	\$210.00	6	1,337.70	North Dakota	Online													
3	AT-007	01-Jan-18	Access	\$227.50	3	668.85	Delaware	Online													
4	AT-012	01-Jan-18	AS9100	\$396.03	3	1,164.31	Oregon	Direct													
5	AT-001	01-Jan-18	Eng Speaking	\$109.64	5	537.22	Missouri	Online													
6	AT-002	01-Jan-18	ISO 14000	\$543.38	6	3,461.30	Colorado	Direct													
7	AT-004	01-Jan-18	ISO 26000	\$284.73	4	1,255.64	Colorado	Direct													
8	AT-003	01-Jan-18	ISO 27001	\$358.75	3	1,054.73	Oregon	Direct													
9	AT-008	01-Jan-18	ISO 31000	\$457.63	3	1,345.42	Rhode Island	Direct													
10	AT-009	01-Jan-18	ISO 50001	\$395.50	3	1,162.77	New Hampshire	Direct													
11	AT-011	01-Jan-18	PowerPoint	\$245.00	3	720.3	South Dakota	Online													
12	AT-010	01-Jan-18	SPSS	\$245.00	3	720.3	Georgia	Online													
13	AT-006	01-Jan-18	TOT	\$280.88	6	1,789.17	Nebraska	Direct													
14	AT-017	02-Jan-18	2D Animation	\$210.00	1	205.8	Ohio	Online													
15	AT-019	02-Jan-18	Access	\$227.50	4	1,003.28	Florida	Online													
16	AT-024	02-Jan-18	AS9100	\$396.03	5	1,940.52	Pennsylvania	Direct													
17	AT-013	02-Jan-18	Eng Speaking	\$109.64	2	268.61	Montana	Online													
18	AT-014	02-Jan-18	ISO 14000	\$543.38	2	1,331.27	Connecticut	Direct													
19	AT-016	02-Jan-18	ISO 26000	\$284.73	2	697.58	Connecticut	Direct													
20	AT-015	02-Jan-18	ISO 27001	\$358.75	2	747.1	Pennsylvania	Direct													
21	AT-020	02-Jan-18	ISO 31000	\$457.63	5	2,242.36	South Carolina	Direct													
22	AT-021	02-Jan-18	ISO 50001	\$395.50	6	2,325.54	New Jersey	Direct													
23	AT-023	02-Jan-18	PowerPoint	\$245.00	5	1,200.50	Tennessee	Online													
24	AT-022	02-Jan-18	SPSS	\$245.00	5	1,200.50	Hawaii	Online													
25	AT-018	02-Jan-18	TOT	\$280.88	1	275.26	Nevada	Direct													
26	AT-029	03-Jan-18	2D Animation	\$210.00	2	514.5	Oklahoma	Online													
27	AT-031	03-Jan-18	Access	\$227.50	2	557.38	Georgia	Online													
28	AT-036	03-Jan-18	AS9100	\$396.03	2	970.26	Rhode Island	Direct													
29	AT-025	03-Jan-18	Eng Speaking	\$109.64	2	268.61	Nebraska	Online													
30	AT-026	03-Jan-18	ISO 14000	\$543.38	2	1,331.27	Delaware	Direct													

General Description

The Excel file for training data management is a powerful tool designed to help you organize, analyze, and track important information related to your organization's training programs. With columns for Training ID, Name, Date, Package Price, Participants, Revenue, Training Location, and Training Type, this file provides a comprehensive solution for managing your training data.

The Training ID column serves as a unique identifier for each training session, making it easy to differentiate and manage multiple training events. The Name column allows you to input the names of your training programs, providing a clear reference for each session. The Date column lets you record the date on which each training session took place, allowing you to track training schedules and timelines.

The Package Price column provides space for recording the cost of each training package, enabling you to track training expenses and revenue generation. The Participants column allows you to input the number of participants who attended each training session, helping you analyze attendance trends and evaluate training program effectiveness. The Revenue column automatically calculates the total revenue generated based on the package price and number of participants, providing you with accurate financial information.

The Training Location column lets you input the location where each training session took place, making it easy to manage and track training venues. The Training Type column provides space for recording the type of training session, such as a workshop, seminar, webinar, or any other type of training, helping you analyze the effectiveness of different training formats.

By using this Excel file for training data management, you can efficiently organize and analyze your training data, gain insights, and make informed decisions to optimize your training programs and improve your organization's overall training effectiveness.

Specific Requirements

- Training ID (Column A): This column should contain unique identifiers for each training session, such as alphanumeric codes or sequential numbers, to easily differentiate and track different training events.
- Name (Column B): This column should allow you to input the names of your training programs or sessions for easy reference.
- Date (Column C): This column should allow you to input the date on which each training session took place, using the appropriate date format.
- Package Price (Column D): This column should allow you to input the cost of each training package, using numerical values or currency format, to track training expenses and revenue generation.
- Participants (Column E): This column should allow you to input the number of participants who attended each training session, using numerical values, to analyze attendance trends and evaluate training program effectiveness.
- Revenue (Column F): This column should automatically calculate the total revenue generated for each training session, based on the package price and number of participants. The formula for calculating revenue could be: $\text{=Package_Price} * \text{Participants}$.
- Training Location (Column G): This column should allow you to input the location where each training session took place, such as the city or venue name, for easy management and tracking.
- Training Type (Column H): This column should allow you to input the type of training session, such as workshop, seminar, webinar, etc., for analysis of different training formats.
- Data Validation: Use data validation to ensure that the input in certain columns, such as Training ID and Participants, meets specific criteria, such as being a unique value or a positive integer.
- Conditional Formatting: Utilize conditional formatting to highlight specific data, such as revenue exceeding a certain threshold or upcoming training sessions based on the date, for visual analysis and identification.
- Sorting and Filtering: Utilize sorting and filtering functions to easily sort and filter data based on different columns, such as sorting by date or filtering by training type, to facilitate data analysis.

- **Pivot Tables:** Utilize pivot tables to summarize and analyze data, such as calculating average revenue per training type or total revenue per month, for comprehensive data analysis.
- R Environment Setup:** Install and configure R on your computer or server, including installing the necessary packages and libraries for your specific data analysis tasks.
- **Data Import and Manipulation:** Utilize R functions and packages to import and manipulate data from various sources, such as CSV files, Excel files, databases, or APIs. Common functions for data import and manipulation include `read.csv()`, `read_excel()`, `dplyr`, and `data.table` packages.
- **Data Cleaning and Preprocessing:** Use R functions and packages to clean and preprocess data, including handling missing values, removing duplicates, converting data types, and handling outliers. Common functions for data cleaning and preprocessing include `na.omit()`, `deduplicated()`, `dplyr`, and `tidyr` packages.
- **Data Visualization:** Utilize R functions and packages to create visualizations, such as bar charts, scatter plots, and heatmaps, to explore and analyze data. Common packages for data visualization include `ggplot2`, `plotly`, and `shiny`.
- **Statistical Analysis:** Utilize R functions and packages to perform statistical analysis on data, such as hypothesis testing, regression analysis, and ANOVA. Common packages for statistical analysis include `stats`, `dplyr`, and `tidyr`.
- **Machine Learning:** Utilize R packages for machine learning, such as `caret`, `randomForest`, and `xgboost`, to build predictive models, perform classification or regression tasks, and evaluate model performance using metrics such as accuracy, precision, recall, and F1-score.
- **Data Transformation and Aggregation:** Utilize R functions and packages to transform and aggregate data, such as reshaping data, merging data frames, and calculating summary statistics. Common functions for data transformation and aggregation include `reshape2`, `dplyr`, and `data.table` packages.
- **Control Flow and Looping:** Utilize R control flow statements, such as `if-else`, `for` loop, and `while` loop, to implement conditional logic and iterative tasks in your R code.
- **Functions and Customization:** Define and utilize custom functions in R to encapsulate repetitive tasks or complex calculations, making your code more modular and reusable.
- **Error Handling:** Implement error handling techniques in your R code, such as `tryCatch()`, to handle exceptions and errors that may occur during data analysis or processing.
- **Reporting and Visualization:** Utilize R markdown and Shiny to create interactive reports and dashboards, incorporating visualizations and data analysis results for presentation and communication purposes.

- Version Control: Use version control tools, such as Git, to track changes in your R code, collaborate with others, and manage code revisions effectively.
- Performance Optimization: Utilize R profiling and benchmarking techniques to identify and optimize performance bottlenecks in your code, ensuring efficient and fast data processing.
- By incorporating these specific requirements, functions, and techniques into your R programming workflow, you can effectively analyze, manipulate, and visualize data, build predictive models, and generate valuable insights for decision-making and data-driven solutions.
- Data Formulas: Utilize Excel formulas, such as SUM, COUNT, AVERAGE, MAX, MIN, etc., to perform various calculations on the data, such as calculating total revenue, average participants, and other key metrics.
- Charts and Graphs: Utilize charts and graphs to visually represent data, such as bar charts, pie charts, or line charts, to provide visual insights and make data analysis more accessible.
- By incorporating these specific requirements, functions, and formulas into your Excel file for training data management, you can effectively organize, analyze, and track your training data, gain valuable insights, and make informed decisions to optimize your training programs and improve overall training effectiveness.

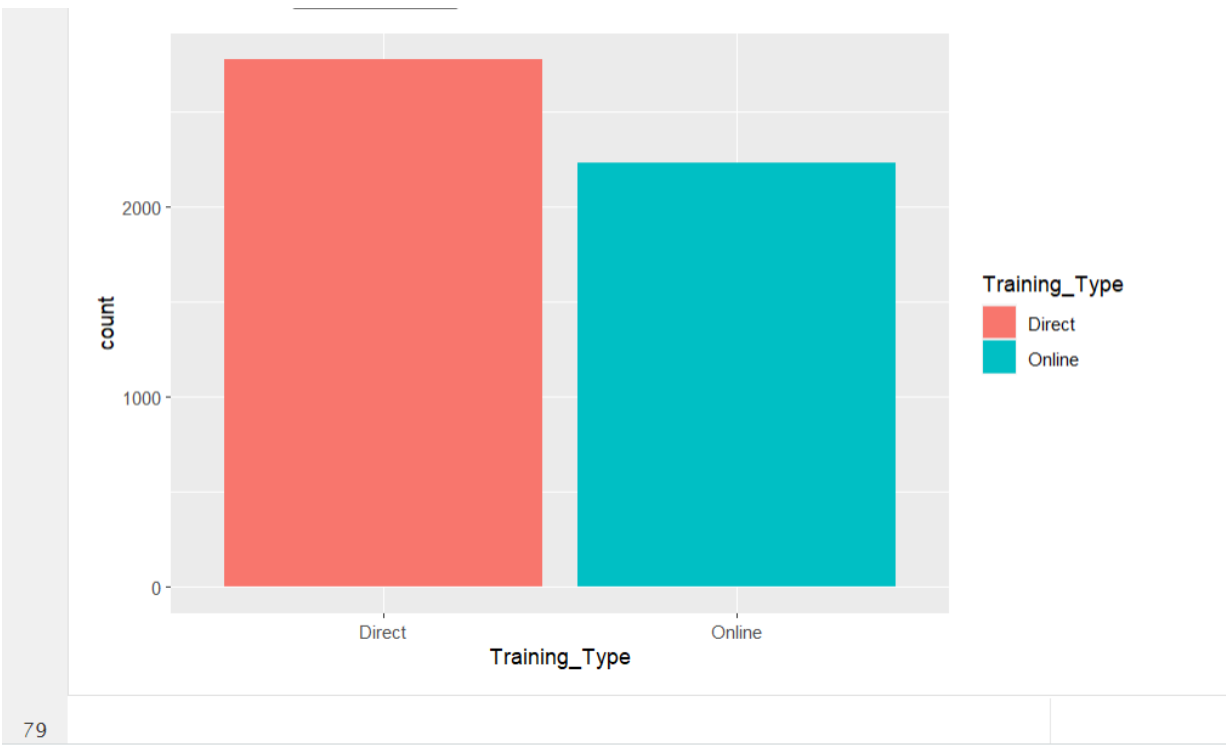
Analysis Result

Analysis results for various parameters related to training and revenue can provide valuable insights into the performance and trends of a training program. Here are some possible analysis results based on the parameters you mentioned:

- 1. **Training Type Wise Number Of Students:** This analysis result provides a breakdown of the number of students enrolled in each training program. It can help identify which training programs are more popular or in demand, and which ones may need improvement. For example:

Direct: 2771 students

Online: 2228 student



2. **Course Wise Number Of Participants:** This analysis result provides a breakdown of the number of participants in each course within a training program. It can help identify which courses are more popular or have higher enrollment rates. For example:

2D Animation: 250 participants

Access: 250 participants

Autodesk: 240 participants

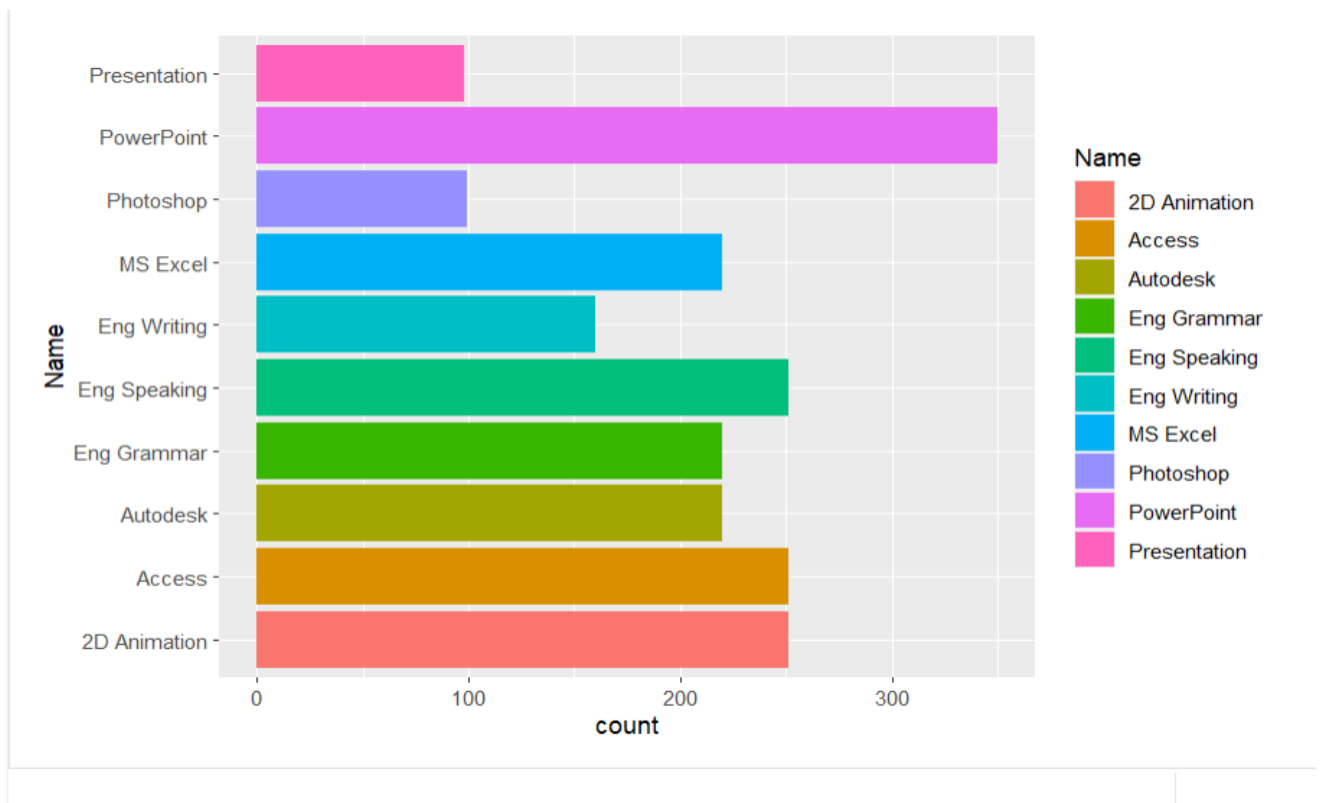
Eng Grammar: 240 participants

Eng Writing: 170 participants

Photoshop: 100 participants

PowerPoint: 350 participants

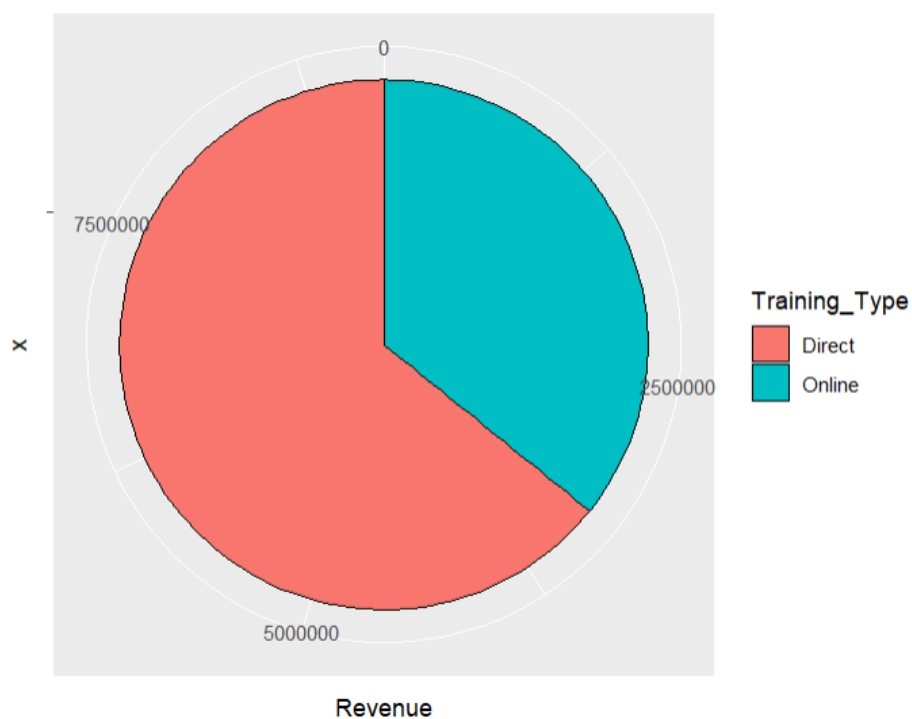
Presentation: 95 participants



3. **Training Revenue Type Wise:** This analysis result provides a breakdown of the revenue generated from different types of training revenue streams, such as registration fees, course fees, certification fees, etc. It can help identify which revenue streams are contributing more to the overall training revenue and can aid in financial planning. For example:

Direct: \$5897525

Online: \$3287386



4. **Revenue Of 10 Different Technology:** This analysis result provides a breakdown of the revenue generated from training programs related to different technologies. It can help identify which technologies are driving higher revenue and which ones may need more focus. For example:

2D Animation: \$530792.2

Access: \$769304.1

Autodesk: \$218294.2

Eng Grammar: \$166094

Eng Speaking: \$290905.2

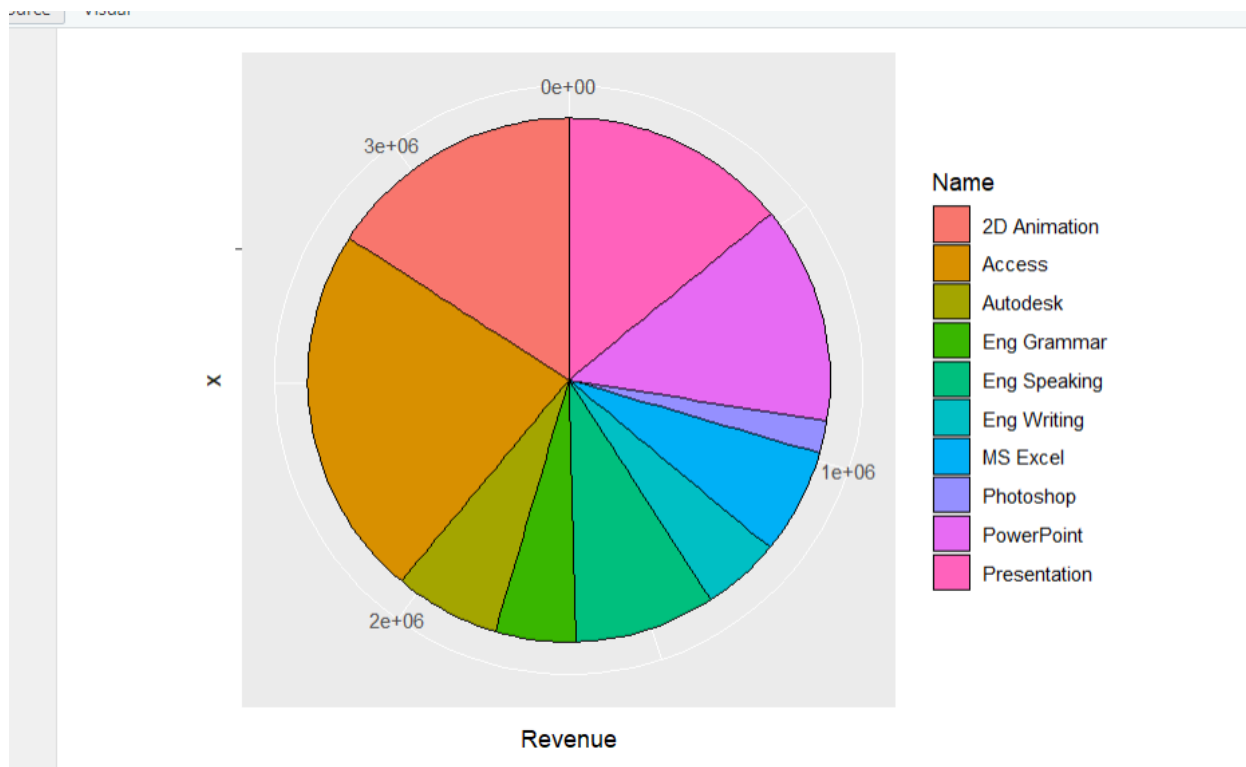
Eng Writing: \$160960.9

MS Excel: \$218087

Photoshop: \$66434.3

PowerPoint: \$449617.9

Presentation: \$469874.7



5. **Number Of Participants Month Wise:** This analysis result provides a breakdown of the number of participants enrolled in training programs on a monthly basis. It can help identify any seasonal trends or patterns in enrollment, which can aid in planning training schedules and resources. For example:

January: 387 participants

February: 436 participants

March: 500 participants

April: 520 participants

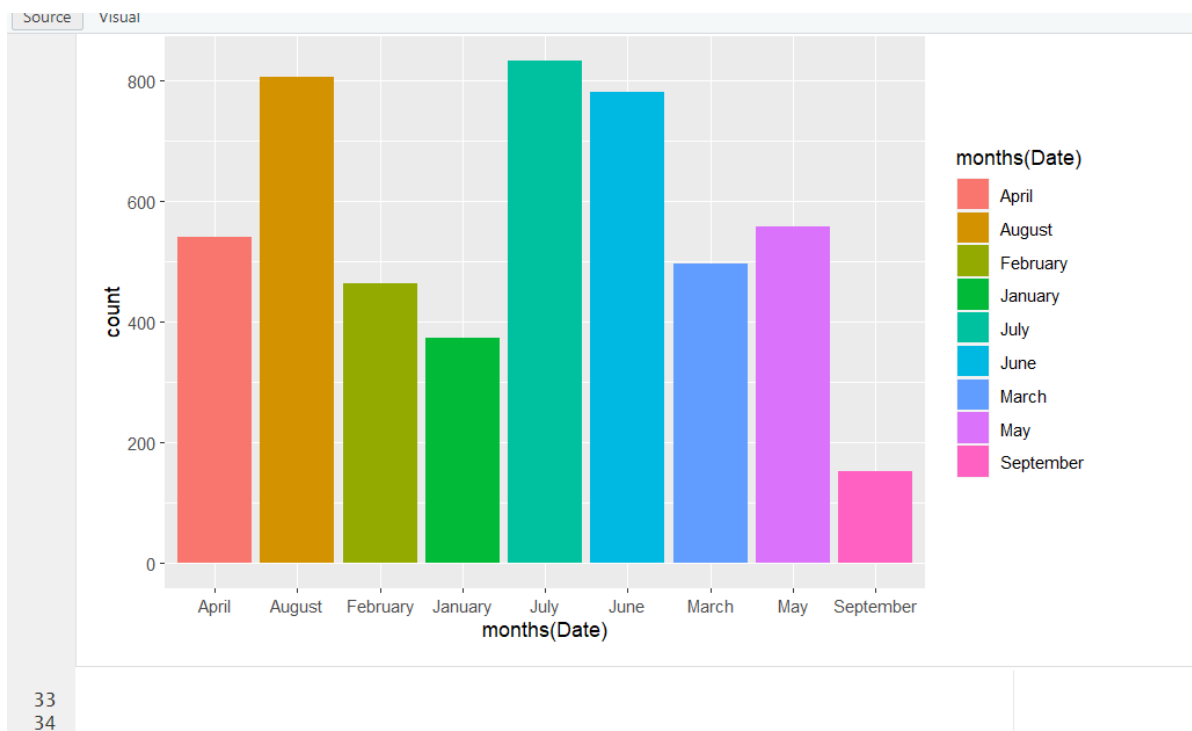
May: 570 participants

June: 780 participants

July: 820 participants

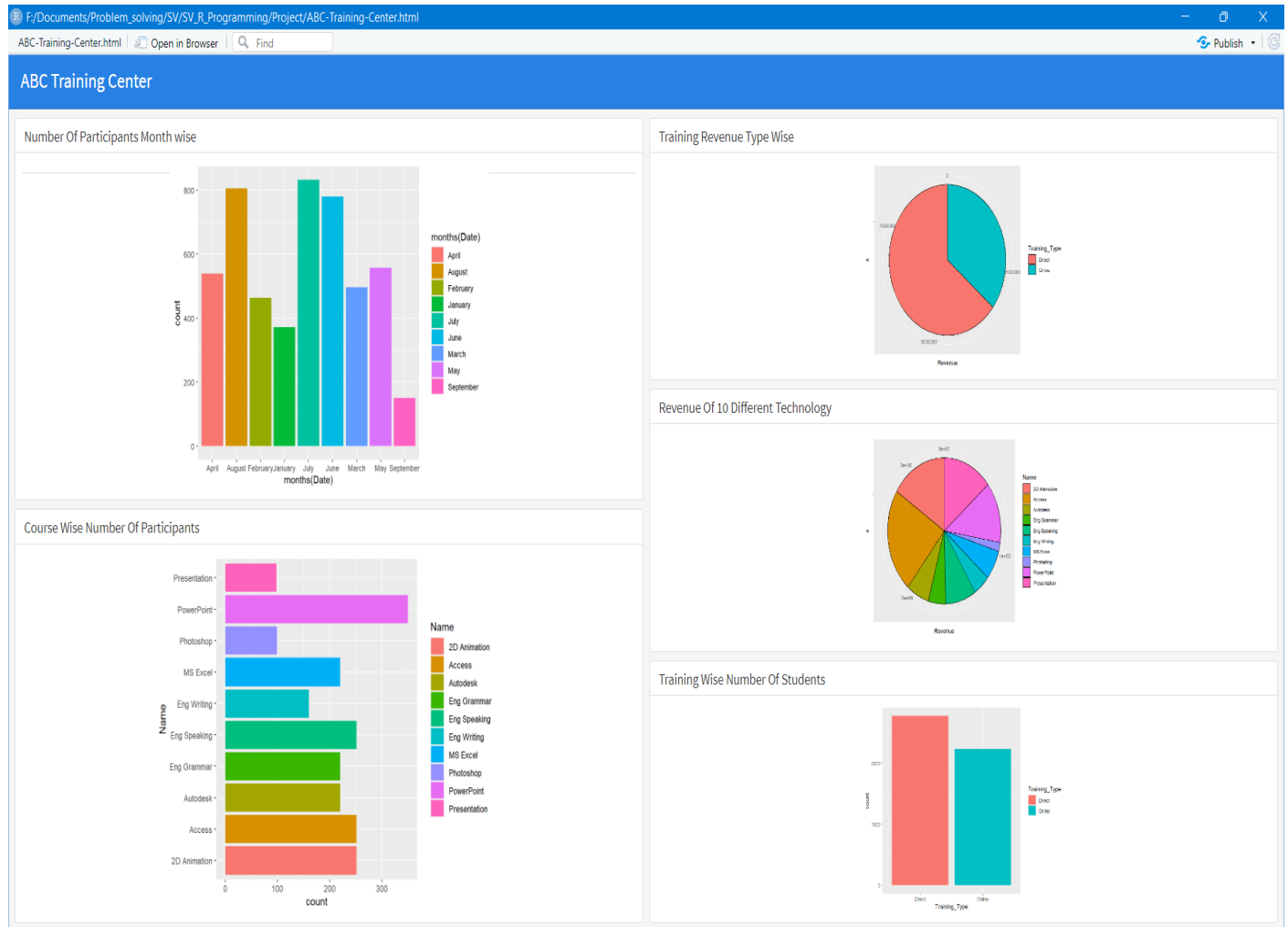
August: 804 participants

September: 130 participants



These analysis results can provide valuable insights into the performance and trends of training programs, which can be used for strategic decision-making, resource allocation, and overall program improvement.

Visualization



List Of Analysis With Result

- The number of students in Direct is 2771 students and Online is 2228 students and the difference is 543 which means in direct the number of students is more.
- The number of participants in PowerPoint is the maximum, which is 350, and in Presentation is minimum which is 95 among 10 different technologies.
- The total training revenue in direct is \$5897525 and online is \$3287386 which means the direct type of revenue is more than online which is \$2610139.
- The total training revenue in Access is the maximum, which is \$769304.1, and Photoshop is \$66434.3 which is minimum from among 10 different technologies.
- The number of participants in July is the maximum, which is 820, and in September is minimum which is 130 from January to September.

Future Scope

- **Advancements in Data Science and Machine Learning:** R has a rich ecosystem of packages for data science and machine learning, and it is likely to continue evolving with new methods, algorithms, and tools for advanced analytics, predictive modeling, and artificial intelligence. Future developments in areas such as deep learning, natural language processing, and reinforcement learning may be incorporated into R's libraries, making it even more powerful for data scientists and machine learning practitioners.
- **Big Data and High-performance Computing:** As data continues to grow in size and complexity, R may see further advancements in big data processing capabilities. New techniques for handling large datasets, distributed computing, and parallel processing may be integrated into R, enabling more efficient and scalable data analysis on big data platforms.
- **Interactive and Dynamic Visualizations:** R has a strong ecosystem of packages for data visualization, and there may be future developments in interactive and dynamic visualizations using advanced graphics capabilities. This may include enhancements in interactive visualization libraries like Shiny, which allows for the creation of web-based interactive applications, making R more accessible and user-friendly for non-programmers.
- **Integration with Cloud Computing and Deployment:** Cloud computing has become increasingly popular for data analysis and machine learning workflows, and R may see further integration with cloud-based tools and services. This may include seamless integration with cloud-based data storage and processing platforms, as well as tools for deploying R-based applications or models on cloud-based environments for scalability and ease of deployment.
- **Collaborative and Open-Source Development:** R has a vibrant community of users and developers who actively contribute to the development of R packages and libraries. The future scope of R may involve more collaborative and open-source development

practices, with continued contributions from the community, leading to the development of new functionalities, bug fixes, and improvements in R's overall ecosystem.

- **Education and Training:** The demand for skilled R programmers and data analysts is expected to continue to grow in the future. Therefore, there may be an increased focus on education and training programs for R programming, both in academic settings and professional training courses. This may include the development of more online resources, tutorials, and educational materials to support learners in mastering R programming skills.



References

- Wickham, H. (2019). Advanced R. CRC Press.
- Murrell, P. (2005). R graphics. The R Journal, 7(1), 3-22.
- R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing. <https://www.r-project.org/>
- Grolemund, G., & Wickham, H. (2017). R for data science: Import, tidy, transform, visualize, and model data. O'Reilly Media.
- Chambers, J. M. (2008). Software for data analysis: Programming with R. Springer Science & Business Media.

Bibliography

- **Google**
- **Kaggle website**
- **Wickham, H. (2019). Advanced R. CRC Press.**