1. **Learning Objectives**

**1.1 Primary Objectives**

1. To establish interest in industrial/Commercial activities
2. To gain work experience enabling the student to apply what He learnt in college and acquire new skills.

**1.2 Secondary Objectives**

1. MS Excel as a Data Analysis tool
2. SQL Server for Data Analysis
3. Python as a Data Analysis tool using Jupyter Notebook & Pandas
4. Analyze Data from multiple disparate sources
5. Tableau as a Data Analysis tool
6. **WEEKLY OVERVIEW OF OJT ACTIVITIES**

|  |  |  |  |
| --- | --- | --- | --- |
| **Week No** | **DATE** | **DAY** | **NAME OF THE TOPIC/MODULE COMPLETED** |
| **Week No-I** | 05-11-2020 | Thursday | Introduction to Data Analysis |
| 06-11-2020 | Friday | Introduction to MS Excel |
| 07-11-2020 | Saturday | MS Excel Data Analysis |
| **Week No-II** | 08-11-2020 | Sunday | MS Excel as a Data Analysis tool |
| 09-11-2020 | Monday | MS Excel Sort & Filter |
| 10-11-2020 | Tuesday | MS Excel Conditional Formatting |
| 11-11-2020 | Wednesday | MS Excel Formulas & Functions |
| 12-11-2020 | Thursday | MS Excel Pivot Tables Analysis |
| 13-11-2020 | Friday | Benefits of Pivot Tables |
| **Week No-III** | 16-11-2020 | Monday | MS Excel Charts |
| 17-11-2020 | Tuesday | SQL Server Data Analysis |
| 18-11-2020 | Wednesday | Introduction to Python |
| 19-11-2020 | Thursday | Data Types & List in Python |
| 20-11-2020 | Friday | Basic Operations & Data Structure in Python |
| 21-11-2020 | Saturday | Python Used to Analyze Data |
| **Week No-IV** | 22-11-2020 | Sunday | Python Functions |
| 23-11-2020 | Monday | Pandas |
| 24-11-2020 | Tuesday | Installing Pandas |
| 25-11-2020 | Wednesday | Installing Pandas with Miniconda |
| 26-11-2020 | Thursday | Data Analysis using Python Libraries |
| 27-11-2020 | Friday | Introduction to Tableau |
| 28-11-2020 | Saturday | Using Tableau for Data Analysis |
| **Week No-V** | 29-11-2020 | Sunday | Tableau Features |
| 30-11-2020 | Monday | General Introduction of Big Data & Hadoop |

1. **Introduction**

This Year’s OJT was on Data Analysis and to learn the role of Junior Data Associate, we learnt a lot of skills like Data Analysis, Microsoft Excel, SQL Server, Python Pandas, Tableau, Big Data and Hadoop. As all of them were new skills of me they were quite challenging but, in the end, I managed to understand them properly thanks to our excellent lecturer Mr. Shubham Kabre.

On-the-job training (OJT) is training that is delivered while an individual is performing tasks or processes related to their particular occupation. The student typically performs tasks that are essential to their job function with the supervision of a manager, coach or mentor. This type of training is typically used to broaden a student’s skill set and to increase productivity.

On-the-job training is an important topic of human resource management. It helps develop the individual and the prosperous growth of the organization. On the job training is a form of training provided at the workplace. During the training, employees are familiarized with the working environment they will become part of organization. Employees also get a hands-on experience using machinery, equipment, tools, materials, etc. Part of on-the-job training is to face the challenges that occur during the performance of the job. An experienced employee or a manager is executing the role of the mentor who through written or verbal instructions and demonstrations are passing on his/her knowledge and company-specific skills to the new employee. Executing the training on at the job location, rather than the classroom, creates a stress-free environment for the employees. On-the-job training is the most popular method of training not only in the United States but in the most of the developed countries, such as the United Kingdom, China, Russia, etc. Its effectiveness is based on the use of existing workplace tools, machines, documents and equipment, and the knowledge of specialists who are working in this field. On-the-job training is easy to arrange and manage and it simplifies the process of adapting to the new workplace. On-the-job training is highly used for practical tasks. It is inexpensive, and it doesn’t require special equipment that is normally used for a specific job. Upon satisfaction of completion of the training, the employer is expected to retain participants as regular employees.

1. **OJT Discussion**

**Subsection 4.1: How the Objectives were achieved?**

The objectives were achieved by performing the job-related activities in a timely and systematic manner.

* 1. First, an Introduction was given about Data Analysis.
  2. Activities related to troubleshooting of various hardware related issues were performed.
  3. Introduction was given about Remote Connection followed by related activities.
  4. Activities related to installation of various Applications and System Softwares were carried out.
  5. Introduction was given about networking and related activities were carried out.
  6. Introduction was given about the Linux operating system.

**Subsection 4.2: What skills were acquired as per the Job Role QP during the OJT?**

**4.2.1 MS Excel as a Data Analysis tool**

Microsoft excel is the most basic tool used for the purpose of data analysis. Anyone who starts to learn the data analysis, starts with learning MS Excel. It is a Microsoft product and has been in use since long. We will see the most important aspects available in excel through this document.

**i) Conditional Formatting:**

Conditional formatting makes it easy to highlight interesting cells or ranges of cells, emphasize unusual values, and visualize data by using data bars, color scales, and icon sets that correspond to specific variations in the data.

Conditional formatting allows you to automatically apply formatting—such as **colors, icons**, and **data bars**—to one or more cells based on the **cell value.** To do this, you'll need to create a **conditional formatting rule.**

**ii) Conditional formatting presets:**

Excel has several predefined styles—or presets—you can use to quickly apply conditional formatting to your data. They are grouped into three categories:

1. Data Bars are horizontal bars added to each cell, much like a bar graph.
2. Colour Scales change the colour of each cell based on its value. Each colour scale uses a two- or three-color gradient.
3. Icon Sets add a specific icon to each cell based on its value.

You can add, delete and manage the rules present in the conditional formatting section.

There is also an option to select Top N or Bottom N sales where N is any positive number. Using this option, we can select top performing or bottom performing values present in the dataset.

**iii) Sort & Filter:**

Sometimes in order to carry out data analysis of huge data we may need to reorder the data either in ascending or descending order as per nature of data present. For e.g. if we have text data then we can rearrange or reorder the same column either in ascending or descending order. This is called as Sort in excel.

For numbers columns we can sort the column in smallest to largest numbers and vice-versa using Sort option. There is an option to sort the columns at multiple levels. With the help of this, we can rearrange the next column once first column is sorted.

Grouping is a useful Excel feature that gives you control over how the information is displayed. You must sort before you can group.

Filter on the other hand is the way to do filtration on the existing data to focus more on filtered aspect.

**iv) Pivot Table:**

A PivotTable is a powerful tool to calculate, summarize, and analyze data that lets you see comparisons, patterns, and trends in your data. A pivot table is a tool in data summation that is common in many business software. It is utilized to collect the summary of a specific data set in a compressed technique. It is a very useful tool in consolidating a large quantity of data that is contained in Microsoft Excel. They let the user make a faster organization and drawing of conclusions from data being collected. Pivot table consists of columns, rows, pages, and data fields. These can be moved around and it assists in expanding, isolating, summarizing, and grouping the specific data. And all of these can be accomplished in real time.

**v) How to create a PivotTable?**

1. Select the cells you want to create a PivotTable from.
   1. (Note: Your data shouldn't have any empty rows or columns. It must have only a single-row heading).
2. Select Insert > PivotTable.
3. Under Choose the data that you want to analyze, Select a table or range.
4. In Table/Range, verify the cell range.
5. Under Choose where you want the PivotTable report to be placed, select New worksheet to place the PivotTable in a new worksheet or Existing worksheet and then select the location you want the PivotTable to appear & select OK.

You can drag any dimensions (text, date etc.) or measures (sales, profit etc.) present in the data in the rows or columns sections in the pivot table as per the need for analysis. On top of pivot table, we can create various charts e.g. bar chart, pie chart etc. to enhance the analysis further.

**vi) Benefits:**

1. User friendly.
2. Creates instant data for analysis.
3. Makes data analysis easier.
4. It summarizes the data easily.
5. Assists in finding patterns in data.
6. Creates accurate reports faster.
7. Helps in making decisions better and faster.

**vii) Formulas & Functions:**

A formula is an expression which calculates the value of a cell. Functions are predefined formulas and are already available in Excel.

Excel uses a default order in which calculations occur. If a part of the formula is in parentheses, that part will be calculated first. It then performs multiplication or division calculations. Once this is complete, Excel will add and subtract the remainder of your formula.

e.g. A1\*A2 + A3, here multiplication will be performed first then addition will be computed.

**viii) Insert Function:**

Every function has the same structure. For example, SUM (A1:A4). The name of this function is SUM. The part between the brackets (arguments) means we give Excel the range A1:A4 as input. This function adds the values in cells A1, A2, A3 and A4. It's not easy to remember which function and which arguments to use for each task. Fortunately, the Insert Function feature in Excel helps you with this.

To insert a function, execute the following steps.

1. Select a cell.
2. Click the Insert Function button.
3. The 'Insert Function' dialog box appears. Search for a function or select a function from a category. For example, choose COUNTIF from the Statistical category.
4. Click OK. The 'Function Arguments' dialog box appears.
5. Click in the Range box and select the range A1:C2.
6. Click in the Criteria box and type >5.
7. Click OK. The function will look like this =COUNTIF (A1:C2,">5")
8. Finally, you will see the desired result.

**ix) Top 10 Functions:**

1. **COUNT:** To count the number of cells that contain numbers, use the COUNT function.
2. **SUM:** To sum a range of cells, use the SUM function.
3. **IF:** The IF function checks whether a condition is met and returns one value if true and another value if false.
4. **AVERAGE:** To calculate the average of a group of numbers, use the AVERAGE function.
5. **COUNTIF:** The COUNTIF function counts the values given within IF clause.
6. **SUMIF:** Sums the values based on IF condition.
7. **VLOOKUP:** Used when the data/value is required to be searched in other sheets.
8. **INDEX and MATCH:** Used to perform advanced lookups.
9. **MIN and MAX:** Used to find minimum and maximum value respectively.
10. SUMPRODUCT: To calculate the sum of the products of corresponding numbers in one or more ranges, use Excel's powerful SUMPRODUCT

**4.2.2 SQL Server for Data Analysis**

For many, SQL is the "meat and potatoes" of data analysis—it's used for accessing, cleaning, and analysing data that's stored in databases. It's very easy to learn, yet it's employed by the world's largest companies to solve incredibly challenging problems.

**i) What is SQL?**

SQL (Structured Query Language) is a programming language designed for managing data in a relational database. It's been around since the 1970s and is the most common method of accessing data in databases today. SQL has a variety of functions that allow its users to read, manipulate, and change data. Though SQL is commonly used by engineers in software development, it's also popular with data analysts for a few reasons:

1. It's semantically easy to understand and learn.
2. Because it can be used to access large amounts of data directly where it is stored, analysts don't have to copy data into other applications.
3. Compared to spreadsheet tools, data analysis done in SQL is easy to audit and replicate. For analysts, this means no more looking for the [cell with the typo in the formula](http://www.washingtonpost.com/blogs/wonkblog/wp/2013/04/16/is-the-best-evidence-for-austerity-based-on-an-excel-spreadsheet-error/).

SQL is great for performing the types of aggregations that you might normally do in an Excel pivot table—sums, counts, minimums and maximums, etc.—but over much larger datasets and on multiple tables at the same time.

**ii) What is Database?**

Database is the repository or storage tank where all the data are dumped or stored in the form of tables which essentially consist of many columns and rows.

**iii) RDBMS**:

RDBMS stands for Relational Database Management System. RDBMS is the basis for SQL, and for all modern database systems like MS SQL Server, IBM DB2, Oracle, MySQL, and Microsoft Access.

A Relational database management system (RDBMS) is a database management system (DBMS) that is based on the relational model as introduced by E. F. Codd.

The data in an RDBMS is stored in database objects which are called as tables. This table is basically a collection of related data entries and it consists of numerous columns and rows.

Remember, a table is the most common and simplest form of data storage in a relational database.

**iv) SQL Constraints:**

Constraints are the rules enforced on data columns on a table. These are used to limit the type of data that can go into a table. This ensures the accuracy and reliability of the data in the database.

Constraints can either be column level or table level. Column level constraints are applied only to one column whereas, table level constraints are applied to the entire table.

Following are some of the most used constraints available in SQL −

1. [**NOT NULL Constraint**](https://www.tutorialspoint.com/sql/sql-not-null.htm)**−** Ensures that a column cannot have a NULL value.
2. [**DEFAULT Constraint**](https://www.tutorialspoint.com/sql/sql-default.htm) **−** Provides a default value for a column when none is specified.
3. [**UNIQUE Constraint**](https://www.tutorialspoint.com/sql/sql-unique.htm)**−** Ensures that all the values in a column are different.
4. [**PRIMARY Key**](https://www.tutorialspoint.com/sql/sql-primary-key.htm)**−** Uniquely identifies each row/record in a database table.
5. [**FOREIGN Key**](https://www.tutorialspoint.com/sql/sql-foreign-key.htm)**−** Uniquely identifies a row/record in any another database table.
6. [**CHECK Constraint**](https://www.tutorialspoint.com/sql/sql-check.htm)**−** The CHECK constraint ensures that all values in a column satisfy certain conditions.
7. [**INDEX**](https://www.tutorialspoint.com/sql/sql-index.htm)**−** Used to create and retrieve data from the database very quickly.

Data Normalization is the process of efficiently organizing the data in the database. There are 2 reasons for this.

1. Eliminating redundant data, e.g. storing the same data in more than one table.
2. Ensuring data dependencies make sense.

**v) SQL DDL, DQL, DML, DCL, AND TCL Commands:**

SQL uses certain commands like Create, Drop, Insert etc. to carry out the required tasks. These SQL commands are mainly categorized into four categories as:

1. **DDL –** Data Definition Language
2. **DQL –** Data Query Language
3. **DML –** Data Manipulation Language
4. **DCL –** Data Control Language

**a) DDL (Data Definition Language):**

DDL or Data Definition Language consists of the SQL commands that can be used to define the database schema. It simply deals with descriptions of the database schema and is used to create and modify the structure of database objects in the database.

e.g. CREATE, DROP, ALTER, TRUNCATE, COMMENT, RENAME

**b) DQL (Data Query Language):**

DML statements are used for performing queries on the data within schema objects. The purpose of DQL Command is to get some schema relation based on the query passed to it.

e.g. SELECT

**c) DML (Data Manipulation Language):**

The SQL commands that deals with the manipulation of data present in the database belong to DML or Data Manipulation Language and this includes most of the SQL statements.

e.g. INSERT, UPDATE, DELETE

**d) DCL (Data Control Language):**

DCL includes commands mainly deals with the rights, permissions and other controls of the database system.

e.g. GRANT, REVOKE

**e) TCL (Transaction Control Language):**

TCL commands deals with the [transaction within the database](https://www.geeksforgeeks.org/sql-transactions/).

e.g. COMMIT, ROLLBACK, SAVEPOINT, SET TRANSACTION

**vi) SQL Functions:**

SQL has different functions for three of its main datatypes i.e. String, Numerical and Date.

**1) String Functions:**

1. **CONCAT –** Adds two or more strings together
2. **CONCAT\_WS -** Adds two or more strings together with a separator
3. **DATALENGTH -** Returns the number of bytes used to represent an expression
4. **FORMAT -** Formats a value with the specified format
5. **LEFT -** Extracts a number of characters from a string (starting from left)
6. **RIGHT -** Extracts a number of characters from a string (starting from right)
7. **LEN -** Returns the length of a string
8. **LTRIM -** Removes leading spaces from a string
9. **RTRIM -** Removes trailing spaces from a string
10. **TRIM -** Removes leading and trailing spaces (or other specified characters) from a string
11. **PATINDEX -** Returns the position of a pattern in a string
12. **REPLACE -** Replaces all occurrences of a substring within a string, with a new substring
13. **REVERSE -** Reverses a string and returns the result
14. **STR -** Returns a number as string
15. **SUBSTRING -** Extracts some characters from a string

**2) Numerical/Math Functions:**

1. **COUNT -** Returns the number of records returned by a select query
2. **MAX –** Returns the maximum value in a set of values
3. **MIN -** Returns the minimum value in a set of values
4. **ROUND -** Rounds a number to a specified number of decimal places
5. **SUM -** Calculates the sum of a set of values

**3) Date Functions:**

1. **DATEADD -** Adds a time/date interval to a date and then returns the date
2. **DATEDIFF -** Returns the difference between two dates
3. **DATEFROMPARTS -** Returns a date from the specified parts (year, month, and day values)
4. **DATENAME -** Returns a specified part of a date (as string)
5. **DATEPART -** Returns a specified part of a date (as integer)
6. **DAY -** Returns the day of the month for a specified date
7. **GETDATE -** Returns the current database system date and time
8. **ISDATE -** Checks an expression and returns 1 if it is a valid date, otherwise 0
9. **MONTH -** Returns the month part for a specified date (a number from 1 to 12)
10. **SYSDATETIME -** Returns the date and time of the SQL Server
11. **YEAR -** Returns the year part for a specified date

**4) Advanced Functions:**

1. **CAST -** Converts a value (of any type) into a specified datatype
2. **COALESCE -** Returns the first non-null value in a list
3. **CONVERT -** Converts a value (of any type) into a specified datatype
4. **CURRENT\_USER -** Returns the name of the current user in the SQL Server database
5. **IIF -** Returns a value if a condition is TRUE, or another value if a condition is FALSE
6. **ISNULL -** Return a specified value if the expression is NULL, otherwise return the expression

**4.2.3 Python as a Data Analysis Tool using Jupyter Notebook & Pandas**

1. First python program as a data analysis tool

* 1. Open jupyter notebook from the start menu.
  2. Create a new notebook i.e. Python 3
  3. Write your first Python program.

2. Python Identifiers

1. An identifier is a name given to entities like class, functions, variables, etc.
2. It helps to differentiate one entity from another.
3. e.g. a=10, name=‘Python’, c=2.4

Python is an increasingly popular tool for data analysis. In recent years, a number of libraries have reached maturity, allowing R and Stata users to take advantage of the beauty, flexibility, and performance of Python without sacrificing the functionality these older programs have accumulated over the years.

**i) Pandas:**

Python itself does not include vectors, matrices, or data frames as fundamental data types. As Python became an increasingly popular language, however, it was quickly realized that this was a major short-coming, and new library were created that added these data-types (and did so in a very, very high-performance manner) to Python. The original library that added vectors and matrices to Python was called numpy. But numpy, while very powerful, was a no-frills library. You couldn’t do things like mix data-types, label your columns, etc.. To remedy this shortcoming a new library was created – built on top of numpy – that added all the nice features we’ve come to expect from modern languages: pandas. Pandas provides a rich set of functions to process various types of data. Further, working with Pandas is fast, easy and more expressive than other tools. Pandas provides fast data

processing as Numpy along with flexible data manipulation techniques as spreadsheets and relational databases. Lastly, pandas integrate well with matplotlib library, which makes it a very handy tool for analysing the data.

**ii) Install Pandas using Anacondas:**

Anaconda is open-source software that contains Jupyter, Spyder, etc. that are used for large data processing, data analytics, heavy scientific computing. If your system is not pre-equipped with Anaconda Navigator.

learn **how to install Anaconda Navigator on**[Windows](https://www.geeksforgeeks.org/how-to-install-anaconda-on-windows/)**or**[Linux](https://www.geeksforgeeks.org/how-to-install-anaconda-on-linux/)**?**   
**iii) Steps to Install Pandas using Anaconda Navigator:**

**Step 1:** Search for **Anaconda Navigator** in Start Menu and open it.

**Step 2:** Click on the **Environment tab** and then click on the **create** button to create a new Pandas Environment.

**Step 3:** Give a name to your Environment, e.g. Pandas and then choose a python version to run in the environment. Now click on the **Create** button to create Pandas Environment.

**Step 4:**Now click on the **Pandas Environment** created to activate it.

**Step 5:** In the list above package names, select **All**to filter all the packages.

**Step 6:** Now in the Search Bar, look for **‘Pandas’**. Select the **Pandas package** for Installation.

**Step 7:** Now Right Click on the checkbox given before the name of the package and then go to **‘Mark for specific version installation**’. Now select the version that you want to install.

**Step 8:** Click on the **Apply**button to install the Pandas Package.

**Step 9:** Finish the Installation process by clicking on the**Apply** button.

**Step 10:** Now to open the Pandas Environment, click on the **Green Arrow** on the right of package name and select the Console with which you want to begin your Pandas programming.

**4.2.4 Analyse Data from multiple disparate sources**

1. Data Analysis is a process of inspecting, cleaning, transforming and modeling data with the goal of discovering useful information, suggesting conclusions and supporting decision-making.
2. It is a technique by the help of which almost all the businesses make growth the market.
3. A well-established business/company is the one which has best data analysis done on its data.

**i) Types:**

1. Data Mining
2. Business Intelligence
3. Statistical Analysis
4. Predictive Analytics
5. Text Analytics

**ii) Data Mining:**

1. Data Mining is the analysis of large quantities of data to extract previously unknown, interesting patterns of data, unusual data and the dependencies.
2. Note that the goal is the extraction of patterns and knowledge from large amounts of data and not the extraction of data itself.
3. Data mining analysis involves computer science methods at the intersection of the artificial intelligence, machine learning, statistics, and database systems.
4. The patterns obtained from data mining can be considered as a summary of the input data that can be used in further analysis or to obtain more accurate prediction results by a decision support system.

**iii) Business Intelligence:**

1. Business Intelligence techniques and tools are for acquisition and transformation of large amounts of unstructured business data to help identify, develop and create new strategic business opportunities.
2. The goal of business intelligence is to allow easy interpretation of large volumes of data to identify new opportunities.
3. It helps in implementing an effective strategy based on insights that can provide businesses with a competitive market-advantage and long-term stability.

**iv) Statistical Analysis:**

1. Statistics is the study of collection, analysis, interpretation, presentation, and organization of data.
2. In data analysis, two main statistical methodologies are used −
   1. **Descriptive statistics −** In descriptive statistics, data from the entire population or a sample is summarized with numerical descriptors.
   2. **Inferential statistics −** It uses patterns in the sample data to draw inferences about the represented population or accounting for randomness.

**v) Predictive Analytics:**

1. Predictive Analytics use statistical models to analyze current and historical data for forecasting (predictions) about future or otherwise unknown events.
2. In business, predictive analytics is used to identify risks and opportunities that aid in decision-making.

**vi) Text Analytics:**

1. Text Analytics, also referred to as Text Mining or as Text Data Mining is the process of deriving high-quality information from text.
2. Text mining usually involves the process of structuring the input text, deriving patterns within the structured data using means such as statistical pattern learning, and finally evaluation and interpretation of the output.

**vii) Data Analysis Tools:**

* 1. Microsoft Excel (SAS/SPSS)
  2. Microsoft SQL
  3. Python/R language
  4. Visualization Tools (Tableau, Power BI etc.)
  5. Pig
  6. Hive
  7. NoSQL/MongoDB
  8. Hadoop/Big Data

**4.2.5 Tableau as a Data Analysis tool**

Tableau is a Business Intelligence tool for visually analyzing the data. Users can create and distribute an interactive and shareable dashboard, which depict the trends, variations, and density of the data in the form of graphs and charts. Tableau can connect to files, relational and Big Data sources to acquire and process data. The software allows data blending and real-time collaboration, which makes it very unique. It is used by businesses, academic researchers, and many government organizations for visual data analysis. It is also positioned as a leader Business Intelligence and Analytics Platform in Gartner Magic Quadrant.

**i) Tableau Sorting & Filtering:**

Sorting of data is a very important feature of data analysis. Tableau allows the sorting of data of the fields, which are called dimensions. There are two ways in which Tableau carries out the sorting.

1. Computed Sorting is the sort directly applied on an axis using the sort dialog button.
2. Manual Sorting is used to rearrange the order of dimension fields by dragging them next to each other in an ad hoc fashion.

**ii) Normal Filters:**

Filtering is the process of removing certain values or range of values from a result set. Tableau filtering feature allows both simple scenarios using field values as well as advanced calculation or context-based filters. In this chapter, you will learn about the basic filters available in Tableau.

There are three types of basic filters available in Tableau. They are as follows −

1. Filter Dimensions are the filters applied on the dimension fields.
2. Filter Measures are the filters applied on the measure fields.
3. Filter Dates are the filters applied on the date fields.

**iii) Condition Filters:**

One of the important filtering options in Tableau is to apply some conditions to already existing filters. These conditions can be very simple like finding only those sales which are higher than a certain amount or it can be a complex one based on a certain formula. The conditions can also be applied to create a range filter.

Top Filters:

The Top option in Tableau filter is used to limit the result set from a filter. For example, from a large set of records on sales you want only the top 10 values. You can apply this filter using the inbuilt options for limiting the records in many ways or by creating a formula. In this chapter, you will explore the inbuilt options.

**iv) Filter Operations:**

Any data analysis and visualization work involves the use of extensive filtering of data. Tableau has a very wide variety of filtering options to address these needs. There are many inbuilt functions for applying filters on the records using both dimensions and measures. The filter option for measures offers numeric calculations and comparison. The filter option for dimension offers choosing string values from a list or using a custom list of values. In this chapter, you will learn about the various options as well as the steps to edit and clear the filters.

Measures are numeric fields. So, the filter options for such fields involve choosing values. Tableau offers the following types of filters for measures.

1. **Range of Values** **−** Specifies the minimum and maximum values of the range to include in the view.
2. **At Least** **−** Includes all values that are greater than or equal to a specified minimum value.
3. **At Most** **−** Includes all values that are less than or equal to a specified maximum value.
4. **Special** **−** Helps you filter on Null values. Include only Null values, Non-null values, or All Values.

Dimensions are descriptive fields having values which are strings. Tableau offers the following types of filters for dimensions.

1. **General Filter** **−** allows to select specific values from a list.
2. **Wildcard Filter** − allows to mention wildcards like cha\* to filter all string values starting with cha.
3. **Condition Filter** **−** applies conditions such as sum of sales.
4. **Top Filter** **−** chooses the records representing a range of top values

**v) Data Source Filters:**

These filters are applied at the level of data pane side where the data has been inserted into Tableau. It is used in case we just want to filter the data on specific dimension/measure before making data available for creating visualizations.

**vi) Formatting:**

We can format any data being pulled in the view as per requirement. We can edit the font, font size, alignment, borders, shading, coloring and can add or remove the grid lines, zero lines etc.

**vii) Calculated Field:**

This is similar to creating functions on top of earlier existing dimensions/measures. E.g. Create expenditure by this formula: SUM(Sales) – SUM(Profit)

We have many functions available that can be used in Calculated Field column. E.g.

Logical, Numerical, String, Date etc.

**viii) Parameters:**

A parameter is a global placeholder value such as a number, date, or string that can replace a constant value in a calculation, filter, or reference line.

For example, you may create a calculated field that returns True if Sales is greater than $500,000 and otherwise returns False. You can replace the constant value of “500000” in the formula with a parameter. Then, using the parameter control, you can dynamically change the threshold in your calculation.

You can even create a *dynamic* parameter that’s set to automatically refresh its current value (to the result of a single-value, view-independent calculation) or list of values (based on a data source column). This will happen each time the workbook is opened and Tableau connects to the data source referenced by the parameter.

**ix) Sets, Groups, Hierarchy & Bins:**

Tableau uses filters, groups and sets to organize data. They were each created for a specific purpose but differentiating between use cases can be difficult at times.

[Set](http://onlinehelp.tableausoftware.com/current/pro/online/en-us/help.htm#sortgroup_sets.html)**:** A subset of your data that meets certain conditions based on existing dimensions.

[Group](http://onlinehelp.tableausoftware.com/current/pro/online/en-us/help.htm#sortgroup_groups.html)**:** Simplifies large numbers of dimension members by combining them into higher-level categories.

Hierarchy: If the requirement is to involve multiple dimensions to be present under one dimension e.g. Country-State-City then hierarchy is used which represents the parent-child relationship.

Bins: Tableau bins are containers of equal size that store data values corresponding to or fitting in bin size. Also, we can say that bins group a set of data into groups of equal interval or size making it a systematic distribution of data. In Tableau, data from any discrete field can be taken to create bins.

**x) Joining the data:**

The data that you analyze in Tableau is often made up of a collection of tables that are related by specific fields (that is, columns). Joining is a method for combining data on based on those common fields. The result of combining data using a join is a virtual table that is typically extended horizontally by adding columns of data.

The data coming from same data source is beneficial if joined through Joins in Tableau.

There are 4 basic types of joins i.e. Inner, Left, Right, Full Outer joins.

**xi) Blending the data:**

When the data from different/heterogenous data sources have to be joined then we use Blending option instead of Join.

**xii) Dashboards:**

A dashboard is a consolidated display of many worksheets and related information in a single place. It is used to compare and monitor a variety of data simultaneously. The different data views are displayed all at once. Dashboards are shown as tabs at the bottom of the workbook and they usually get updated with the most recent data from the data source. While creating a dashboard, you can add views from any worksheet in the workbook along with many supporting objects such as text areas, web pages, and images.

Each view you add to the dashboard is connected to its corresponding worksheet. So when you modify the worksheet, the dashboard is updated and when you modify the view in the dashboard, the worksheet is updated.

You can pull any worksheets in horizontal/vertical sense in the dashboard pane to create visualizations.

There are some objects present in the Tableau desktop to add more aspects in the dashboard.

1. Horizontal
2. Vertical
3. Text
4. Blank
5. Navigation
6. Image
7. Extension
8. Webpage

By default, the orientation in Tableau dashboard is ‘Tiled’ means the Tableau will automatically adjust the width and height of the worksheets being pulled in the view. If we want to adjust the width and height, then instead of Tiled orientation we should go for ‘Floating’ orientation.

**xiii) Actions in Dashboard:**

To make interactive dashboards, Tableau has provided the option of Actions where we can select any action option to make the different worksheets, which are pulled in dashboard, talk each other.

E.g. If in the bar graph we clicked on any bar then automatically with respect to that bar graph value other components in dashboard should get changed. For this purpose, we use Actions.

There are 6 types of actions present in Tableau.

1. Filter
2. Highlight
3. Go to URL
4. Go to Sheet
5. Change Parameters
6. Change Set Values

**xiv) Tableau Stories:**

Use stories to make your case more compelling by showing how facts are connected, and how decisions relate to outcomes. You can then publish your story to the web, or present it to an audience.

Each story point can be based on a different view or dashboard, or the entire story can be based on the same visualization seen at different stages, with different filters and annotations.

If the requirement is to combine everything, we created including dashboard then we go for making stories in Tableau. So, Story in Tableau may consist of worksheets alone or coupled with dashboards as well.

**Subsection 4.3: Results / observations / work / experience / Practical’s done as per the Job Role QP in the OJT Company.**

The following Practical’s were completed successfully during the OJT.

1. Using MS Excel for Data Analysis
2. Using Conditional Formatting in MS Excel
3. Using Functions in MS Excel
4. Using Pivot Table in MS Excel
5. Using SQL Server for Data Analysis
6. Data Structures in Python
7. Functions in Python
8. Basic syntax of Python
9. Using Pandas library in Python
10. Using Tableau for Data Analysis
11. List Operations
12. Tuples

**Subsection 4.4: What challenges were experienced during OJT?**

The following minor challenges were faced during the OJT:

1. Learning new technologies can be tiresome.
2. There was a little difficulty in understanding new terminology.
3. Overwhelming complexity.
4. Resistance to change.
5. **Conclusion**

The Student was able to acquire new skills within the specified job role of Infrastructure Engineer in a timely and structured manner.

The Following Objectives were completed in the training: MS Excel as a Data Analysis tool, SQL Server for Data Analysis, Python as a Data Analysis tool using Jupyter Notebook & Pandas, Analyze Data from multiple disparate sources, Tableau as a Data Analysis tool

1. **Bibliography**
2. Study.com
3. GeeksforGeeks.org
4. StackOverflow.com
5. Developer.mozilla.com
6. W3schools.com
7. Wikipedia.org
8. www.tutorialspoint.com/tableau/
9. Docs.python.org
10. Support.microsoft.com/en-us/excel
11. Pandas.pydata.org/docs