**INTRODUCTION TO DATA MANAGEMENT**

(Project Semester August-December 2020)

***SOLAR AND LUNAR ECLIPSE***

Submitted by

Sunnam Sai Jyothi Swaroop

11802400

K18GE

INT217

Under the Guidance of

**Komal Arora-17783**

**Discipline of CSE/IT**

**Lovely School of Technology and sciences**

**Lovely Professional University, Phagwara**



**CERTIFICATE**

This is to certify that Sunnam Sai Jytohi Swaroop bearing Registration no. 11802400 has completed int217 project titled, **“ Data Management”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

**Komal Arora-17783**

**Lecturer of Introduction to Data Management**

**School of technology and sciences**

Lovely Professional University

Phagwara, Punjab.

**DECLARATION**

I, Sunnam Sai Jyothi Swaroop student of Introduction to data management under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: November 14, 2020 Signature: S.S.J.SWAROOP

Registration No. 11802400 sunnam sai Jyothi swaroop

**Acknowledgement**

I would like to express my special thanks of gratitude to my teacher Komal Arora maam as who gave me the golden opportunity to do this wonderful project on the topic Introduciton to Data Management, which also helped me in doing a lot of Research and i came to know about so many new things I am really thankful to them.Secondly i would also like to thank my parents and friends who helped me a lot in finalizing this project within the limited time frame.

I am are over helmed in all humbleness and gratefulness to acknowledge my depth to all those who have helped me to put these ideas, well above the level of simplicity and into something concrete.  
I would like to express my special thanks of gratitude to my teacher who gave me the golden opportunity to do this wonderful project on the topic "Data Management" , which also helped me in doing a lot of Research and i came to know about so many new things. I am really thankful to them.  
Any attempt at any level can 't be satifactorily completed without the support and guidance of MY parents and friends.  
I would like to thank my parents who helped me a lot in gathering different information, collecting data and guiding me from time to time in making this project , despite of their busy schedules ,they gave me different ideas in making this project unique.  
Thanking you,  
Sunnam Sai Jyothi Swaroop-11802400  
K18GE

**Table of Content**

1. Introduction
2. Objectives/Scope of the Analysis
3. Source of dataset
4. ETL process
5. Analysis on dataset (for each analysis)
6. Introduction
7. General Description
8. Specific Requirements, functions and formulas
9. Analysis results
10. Visualization
11. List of Analysis with results
12. References
13. Bibliography

**Introduction :-**

Data Analysis is a process of inspecting, cleaning, transforming and modeling data with the goal of discovering useful information, suggesting conclusions and supporting decisionmaking.

**Types of Data Analysis**

Several data analysis techniques exist encompassing various domains such as business, science, social science, etc. with a variety of names. The major data analysis approaches are-

• Data Mining

• Business Intelligence

• Statistical Analysis

• Predictive Analytics

• Text Analytics

**Data Mining**

Data Mining is the analysis of large quantities of data to extract previously unknown, interesting patterns of data, unusual data and the dependencies. Note that the goal is the extraction of patterns and knowledge from large amounts of data and not the extraction of data itself.

Data mining analysis involves computer science methods at the intersection of the artificial intelligence, machine learning, statistics, and database systems.

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.

**Business Intelligence**

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. The goal of business intelligence is to allow easy interpretation of large volumes of data to identify new opportunities. It helps in implementing an effective strategy based on insights that can provide businesses with a competitive market-advantage and long-term stability.

**Statistical Analysis**

Statistics is the study of collection, analysis, interpretation, presentation, and organization of data.

In data analysis, two main statistical methodologies are used-

• **Descriptive statistics**: In descriptive statistics, data from the entire population or a sample is summarized with numerical descriptors such as

o Mean, Standard Deviation for Continuous Data

o Frequency, Percentage for Categorical Data

• **Inferential statistics:** It uses patterns in the sample data to draw inferences about the represented population or accounting for randomness. These inferences can beo answering yes/no questions about the data (hypothesis testing)

o estimating numerical characteristics of the data (estimation)

o describing associations within the data (correlation)

o modeling relationships within the data (E.g. regression analysis)

**Predictive Analytics**

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

**Text Analytics**

Text Analytics, also referred to as Text Mining or as Text Data Mining is the process of deriving high-quality information from text. 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.

**Data Analysis Process**

Data Analysis is defined by the statistician John Tukey in 1961 as "Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data.”

Thus, data analysis is a process for obtaining large, unstructured data from various sources and converting it into information that is useful for-

• Answering questions

• Test hypotheses

• Decision-making

• Disproving theories

**Data Analysis with Excel**

Microsoft Excel provides several means and ways to analyze and interpret data. The data can be from various sources. The data can be converted and formatted in several ways. It can be analyzed with the relevant Excel commands, functions and tools - encompassing Conditional Formatting, Ranges, Tables, Text functions, Date functions, Time functions, Financial functions, Subtotals, Quick Analysis, Formula Auditing, Inquire Tool, What-if Analysis, Solvers, Data Model, PowerPivot, PowerView, PowerMap, etc.

You will be learning these data analysis techniques with Excel as part of two parts-

• Data Analysis with Excel and

• Advanced Data Analysis with Excel

Data Analysis is a process of collecting, transforming, cleaning, and modeling data with the goal of discovering the required information. The results so obtained are communicated, suggesting conclusions, and supporting decision-making. Data visualization is at times used to portray the data for the ease of discovering the useful patterns in the data. The terms Data Modeling and Data Analysis mean the same.

Data Analysis Process consists of the following phases that are iterative in nature-

• Data Requirements Specification

• Data Collection

• Data Processing

• Data Cleaning

• Data Analysis

• Communication

The data required for analysis is based on a question or an experiment. Based on the requirements of those directing the analysis, the data necessary as inputs to the analysis is identified (e.g., Population of people). Specific variables regarding a population (e.g., Age and Income) may be specified and obtained. Data may be numerical or categorical.

**Data Collection**

Data Collection is the process of gathering information on targeted variables identified as data requirements. The emphasis is on ensuring accurate and honest collection of data. Data Collection ensures that data gathered is accurate such that the related decisions are valid. Data Collection provides both a baseline to measure and a target to improve.

Data is collected from various sources ranging from organizational databases to the information in web pages. The data thus obtained, may not be structured and may contain irrelevant information. Hence, the collected data is required to be subjected to Data Processing and Data Cleaning.

**Data Processing**

The data that is collected must be processed or organized for analysis. This includes structuring the data as required for the relevant Analysis Tools. For example, the data might have to be placed into rows and columns in a table within a Spreadsheet or Statistical Application. A Data Model might have to be created.

**Data Cleaning**

The processed and organized data may be incomplete, contain duplicates, or contain errors. Data Cleaning is the process of preventing and correcting these errors. There are several types of Data Cleaning that depend on the type of data. For example, while cleaning the financial data, certain totals might be compared against reliable published numbers or defined thresholds. Likewise, quantitative data methods can be used for outlier detection that would be subsequently excluded in analysis.

**Data Analysis**

Data that is processed, organized and cleaned would be ready for the analysis. Various data analysis techniques are available to understand, interpret, and derive conclusions based on the requirements. Data Visualization may also be used to examine the data in graphical format, to obtain additional insight regarding the messages within the data.

Statistical Data Models such as Correlation, Regression Analysis can be used to identify the relations among the data variables. These models that are descriptive of the data are helpful in simplifying analysis and communicate results.

The process might require additional Data Cleaning or additional Data Collection, and hence these activities are iterative in nature.

**Communication**

The results of the data analysis are to be reported in a format as required by the users to support their decisions and further action. The feedback from the users might result in additional analysis.

The data analysts can choose data visualization techniques, such as tables and charts, which help in communicating the message clearly and efficiently to the users. The analysis tools provide facility to highlight the required information with color codes and formatting in tables and charts.

**EXCEL DATA ANAYISIS:-**

Excel provide commands, functions and tools that make your data analysis tasks easy. You can avoid many time consuming and/or complex calculations using Excel. In this tutorial, you will get a head start on how you can perform data analysis with Excel. You will understand with relevant examples, step by step usage of Excel commands and screen shots at every step.

**Ranges and Tables**

The data that you have can be in a range or in a table. Certain operations on data can be performed whether the data is in a range or in a table.

However, there are certain operations that are more effective when data is in tables rather than in ranges. There are also operations that are exclusively for tables.

You will understand the ways of analyzing data in ranges and tables as well. You will understand how to name ranges, use the names and manage the names. The same would apply for names in the tables.

**Data Cleaning – Text Functions, Dates and Times**

You need to clean the data obtained from various sources and structure it before proceeding to data analysis. You will learn how you can clean the data.

* With Text Functions
* Containing Date Values
* Containing Time Values

**Conditional Formatting**

Excel provides you conditional formatting commands that allow you to color the cells or font, have symbols next to values in the cells based on predefined criteria. This helps one in visualizing the prominent values. You will understand the various commands for conditionally formatting the cells.

**Sorting and Filtering**

During the preparation of data analysis and/or to display certain important data, you might have to sort and/or filter your data. You can do the same with the easy to use sorting and filtering options that you have in Excel.

**Subtotals with Ranges**

As you are aware, PivotTable is normally used to summarize data. However, Subtotals with Ranges is another feature provided by Excel that will allow you to group / ungroup data and summarize the data present in ranges with easy steps.

**Quick Analysis**

With Quick Analysis tool in Excel, you can quickly perform various data analysis tasks and make quick visualizations of the results.

**Understanding Lookup Functions**

Excel Lookup Functions enable you to find the data values that match a defined criteria from a huge amount of data.

**PivotTables**

With PivotTables you can summarize the data, prepare reports dynamically by changing the contents of the PivotTable.

**Data Visualization**

You will learn several Data Visualization techniques using Excel Charts. You will also learn how to create Band Chart, Thermometer Chart, Gantt chart, Waterfall Chart, Sparklines and PivotCharts.

**Data Validation**

It might be required that only valid values be entered into certain cells. Otherwise, they may lead to incorrect calculations. With data validation commands, you can easily set up data validation values for a cell, an input message prompting the user on what is expected to be entered in the cell, validate the values entered with the defined criteria and display an error message in case of incorrect entries.

**Financial Analysis**

Excel provides you several financial functions. However, for commonly occurring problems that require financial analysis, you can learn how to use a combination of these functions.

**Working with Multiple Worksheets**

You might have to perform several identical calculations in more than one worksheet. Instead of repeating these calculations in each worksheet, you can do it one worksheet and have it appear in the other selected worksheets as well. You can also summarize the data from the various worksheets into a report worksheet.

**Formula Auditing**

When you use formulas, you might want to check whether the formulas are working as expected. In Excel, Formula Auditing commands help you in tracing the precedent and dependent values and error checking.

**Inquire**

Excel also provides Inquire add-in that enables you compare two workbooks to identify changes, create interactive reports, and view the relationships among workbooks, worksheets, and cells. You can also clean the excessive formatting in a worksheet that makes Excel slow or makes the file size huge.

**Advanced Data Analysis:-**

Excel provides several commands, functions and tools that make your complex data analysis tasks easy. Excel lets you perform various complex calculations with ease. In this tutorial, you will understand the versatile data analysis tools of Excel. You will understand data analysis with relevant examples, step by step instructions and screen shots at every step.

### **Data Consolidation**

You might have to consolidate the data from various sources and present a report. The data could be in the worksheets of the same workbook or in different workbooks. With Excel data tool Consolidate, you can perform this in a few easy steps.

## What-If Analysis

What-If Analysis provides you tools to handle the following data analysis situations −

* Find the input values that result in a specified value. The result could be set up as a formula with the input values as variables. By varying the values of the input variables, Excel provides the solution with the Goal Seek Tool.
* Find the possible output values by varying the values of one or two variables. The result could be set up as a formula with one or two input values as variables. By varying the values for the input variables, Excel provides the solution with the Data Table Tool.
* Find the possible output values that are a result of varying the values of more than two variables. The result could be set up as a formula with the input values as variables. By varying the values for the input variables, Excel provides the solution with the Scenario Manager Tool.

### **Optimizing with Excel Solver Add-in**

Solver is used to handle complex goal seek situations. In such cases, in addition to the inputs and outputs, there will be defined constraints or limits imposed on the possible input values. Further, Solver is used to result in an optimal solution.

Excel has a Solver Add-in that helps you solve such complex problems.

## Importing Data into Excel

Your data analysis might depend on various external data sources. In Excel, you can import data from different data sources, such as Microsoft Access Database, Web Pages, Text Files, SQL Server Table, SQL Server Analysis Cube, XML File, etc.

You can import any number of data tables simultaneously from a database. When you are importing multiple tables from a relational database such as Access, the existing relationships among the tables will be retained in Excel also. While importing the data, you can also optionally create a PivotTable or PivotChart or Power View report based on that data.

You can just create a data connection with a data source, or import the data into Excel. If you import the data into Excel, the data tables are added to the Data Model in Excel.

### **Data Model**

Data Model in Excel is used to integrate data from multiple tables in the current workbook and / or from the imported data and / or from the data sources connected to the workbook through data connections. Data model is used transparently in PivotTable, PivotChart, PowerPivot and Power View reports.

* You can create a Data Model while importing data, or from the Excel tables in the workbook.
* The data tables in the Data Model can be viewed either in Data View or Diagram View.
* With a Data Model, you can create relationships among the data tables.
* You can either use the Create Relationship command or just click and drag and connect the fields in the two tables that define the relationship in the diagram view of the Data Model.

### **Exploring Data with PivotTable**

As you can integrate the Data Model with a PivotTable, you can do extensive data analysis by collating, connecting, summarizing and reporting data from several different sources. As you can import tables from external data sources and create a PivotTable, it is possible to have automatic updates of the values in the PivotTable whenever the data in the connected data sources is updated.

You can create a PivotTable with the fields from multiple tables, provided the tables have relationships defined. If a relationship does not exist, Excel prompts you to create one and you can do so from the PivotTable itself. The relationship that you so define is reflected in the Data Model.

### **Exploring Data with PowerPivot**

You can use PowerPivot to access, analyze and report data from various data sources. PowerPivot can help you handle large data with ease and produce fascinating analysis reports.

PowerPivot provides you commands to manage the Data Model, add Excel tables to Data Model, to add calculated fields in the Data Tables, to define KPIs, etc.

**TABLEAU**

At present the data is been generated everywhere like YouTube, Tumblr, Reddit, Facebook, WhatsApp, Twitter, Instagram Gmail, LinkedIn and Academia. Understanding this data is very important as this is crucial and very important entity of an Organization, Nations, and Institutions. Big data is a collection of large and complex data which are difficult to be handled with traditional data processing application software.

Analyzing and visualizations the data sets can give a new business trends, prevent diseases, and model to forecast future paradigms and combat crime and so on. At this time, the most used tools for data analytics and visualizations, data discovery are tableau .

Tableau is one of the fastest upcoming business intelligence (BI) tool. It is fast to deploy, easy to learn and very useful for a user.

Tableau is a software that can help users explore and understand their data by creating interactive visualizations. The software has the advantages that it can be used in conjunction with almost any database, and it is easy to use by dragging and dropping to create an interactive visualization expressing the desired format.

The COVID-19(Coronavirus) has proven to be a truly global pandemic, impacting people in just about every corner of theworld. The usefulness of Tableau in COVID-19(Coronavirus) data analytics can be measured by its performance, user friendly environment, and better decisions .

Tableau connects users with a variety of data sources and enables them to create data visualizations by making charts, maps, dashboards, and stories through a simple drag and drop interface. This paper presents a procedure for the interactive visualization and analysis of COVID-19(Coronavirus) data using Tableau as a intelligence tool.

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.Data analysis is very fast with Tableau and the visualizations created are in visual form with dashboards and worksheets.

A Tableau dashboard allows for multiple visualizations to be seen within a single view. It’s often used to show only the most important data and is sometimes personalized. It works by connecting to data stored in various places. It can pull data from any source imaginable.

From a simple excel sheet to PDF to complex database like Oracle to the next level of cloud such as Amazon webs services, Microsoft Azure SQL database , Google Cloud SQL, can be extracted by Tableau.Therefore,we are introduces Tableau and presents the procedure ofusing Tableau for the interactive visualization and analysis of COVID-19(Coronavirus) data to encourage its widespread use.Tableau is a new age data analytics and visualizations tool that provides flexibility and ease-of-use with a smooth experience to the users. speed. A tableau, a tool use for complex visualization and simplification of complex data.

It was designed to help the user to create visuals and graphics without the help of any programmer or any prior knowledge of programming. Data visualization is an intuitive way for users to easily read and understand data, especially in big data analyses .

It helps to improve the quality of governance policies or services by presenting an integrated view and evidence for making

**Data Visualization**

Data visualization is the use of human natural skills to enhance data processing and organization efficiency. Data visualization is the representation of data or information in a graph, chart, or other visual format. Visualization can help us deal with more complex information and enhance memory .

It communicates relationships of the data with images. This is important because it allows trends and patterns to be more easily seen. With the rise of big data upon us, we need to be able to interpret increasingly larger batches of data.

Machine learning makes it easier to conduct analyses such as predictive analysis, which can then serve as helpful visualizations to present .

But data visualization is not only important for data scientists and data analysts, it is necessary to understand data visualization in any career. Whether you work in finance, marketing, tech, design, or anything else, we need to visualize data.

That fact showcases the importance of data visualization. Its main goal is to distill large datasets into visual graphics to allow for easy understanding of complex relationships within the data . It is often used interchangeably with terms such as information graphics, statistical graphics, and information visualization.

Data visualization is a huge field with many disciplines. It is precisely because of this interdisciplinary nature that the visualization field is full of vitality and opportunities.

**Objectives/Scope of the Analysis:-**

### **1. Overview sheet**

Excel dashboards make it easy to perform quick overviews of data reports rather than going through large volumes of data. Overviews help in making quick and urgent decisions since one can skim through a lot of information at once and within a short time.

The dashboards help in tracking [Key Performance Indicators (KPIs)](https://corporatefinanceinstitute.com/resources/knowledge/finance/key-performance-indicators-kpis/) with ease, which helps organizations track the progress on their targets. They provide a high-level summary of key aspects of your data to keep everyone at par with the progress, hence giving the organization a timely indicator for necessary action in real-time.

Excel d`ashboards include various elements such as [charts](https://corporatefinanceinstitute.com/resources/excel/study/how-to-make-a-graph-in-excel/), tables, figures, and gauges that help in presenting the data. They can handle any type of data from different market and purposes, and the information can be used for marketing, financial, or other business projects. The dashboard is most applicable to large volumes of data since it would otherwise be hectic to go through such large volumes of data, especially with limited time. One can build their dashboard or use a template to begin using Excel dashboards.

### **Creation of an Excel Dashboard**

The following outlines the initial steps in creating an Excel dashboard:

#### **1. Brainstorm ideas and strategize on the dashboard’s main purpose**

Before investing time and money to build Excel dashboards, users should first brainstorm ideas on the type of data to add to the dashboard. Strategize on the main purpose you want the dashboard to serve. Do you want to track certain departments of the business or the performance of a specific product produced by the company?

#### **2. identify the appropriate data source**

After deciding on the purpose, the next step is to identify the appropriate source of the data that is going to be displayed on the dashboard. The data forms the basic element of the dashboard and guides the components that will be added to it.

The purpose of creating the dashboard determines, to a large extent, its appearance and features. The dashboard should encompass only the necessary aspects of the data that is relevant in making key decisions. Its appearance also depends on the recipients of the information. What are their preferences? Is the consumer a manager, external client, or a colleague? How much time do they have to study the dashboard? All the attributes should be key in designing the dashboard while keeping in mind the consumer’s preferences.

### **Design of an Excel Dashboard**

The brainstorming stage will outline relevant dashboard elements to include in the design. You can decide to use or improve prebuilt templates to save time and money. The key elements of the template will include [pivot tables](https://corporatefinanceinstitute.com/resources/excel/study/pivot-table-guide/), static tables, dynamic charts, auto-shape objects, gauge widgets, and other non-chart widgets.

The space occupied by each of the items also determines the appearance and readability of the dashboard. Are there too many small objects in the dashboard? Are the elements necessary, or do you need a few large objects that are easy and fast to study? Identify all the key elements that you will want to see on the dashboard so that you can categorize similar elements in the same section within the dashboard.

In addition, the Excel dashboard background color affects the readability of the data to a large extent. You can choose to color code similar objects to make it easy for the data users to read the information presented on the dashboard. The choice of colors also helps users distinguish between certain groups of elements for easier comparison. The Excel dashboard’s user interface can be enhanced by simplifying the navigation panels. One way to achieve it is to add labels to graphs, include drop-down lists, and freeze panels to limit scrolls.

### **How PowerPoint Eases Use of Excel Dashboards**

Microsoft’s slideshow and presentation application, [PowerPoint](https://office.live.com/start/PowerPoint.aspx), can help in the presentation of Excel dashboards by making it more interactive for users. The option of making the dashboard interactive using Excel only can be a complex process since it would require using macros and [VBA](https://corporatefinanceinstitute.com/resources/excel/study/vba-excel-example/), which are complex languages to use with Excel sheets. The simpler alternative is to add the interactive elements to PowerPoint since the application comes with in-built elements.

For example, you can add charts to PowerPoint and use the interactivity elements to simplify the presentation. If you are presenting the performance data of ABC Limited for ten years, it will be difficult to present the data in Excel. When using PowerPoint, you can create ten charts for the ten years, with each chart showing the company’s performance for each year.

The charts can then be added to the Excel dashboard. Each chart can be placed on an individual slide, such that when navigating through the charts, the slides will appear as if they are in motion. It increases the interactivity and presentation of the ten-year performance data in Excel.

For those who are new to dashboards, it would be ideal to get an understanding of the dashboards first. In this chapter, you will get to know the definition of dashboard, how it got its name, how they became popular in IT, key metrics, benefits of dashboards, types of dashboards, dashboard data and formats and live data on dashboards.

In information technology, a dashboard is an easy to read, often single page, real-time user interface, showing a graphical presentation of the current status (snapshot) and historical trends of an organization’s or department’s key performance indicators to enable instantaneous and informed decisions to be made at a glance.

Dashboards take their name from **automobile dashboards**. Under the hood of your vehicle, there may be hundreds of processes that impact the performance of your vehicle. Your dashboard summarizes these events using visualizations so that you have the peace of mind to concentrate on safely operating your vehicle. In a similar way, business dashboards are used to view and/or monitor the organization’s performance with ease.

The idea of **digital dashboards** emerged from the study of decision support systems in the 1970s. Business dashboards were first developed in the 1980s, but due to the problems with data refreshing and handling, they were put on the shelf. In the 1990s, the information age quickened pace and data warehousing, and online analytical processing (OLAP) allowed dashboards to function adequately. However, the use of dashboards did not become popular until the rise of key performance indicators (KPIs), and the introduction of Robert S. Kaplan and David P. Norton's Balanced Scorecard. Today, the use of dashboards forms an important part of decision making.



In today’s business environment, the tendency is towards Big Data. Managing and extracting real value from all that data is the key for modern business success. A welldesigned dashboard is a remarkable information management tool.

## Dashboard – Definition

Stephen Few has defined a dashboard as “a visual display of the most important information needed to achieve one or more objectives which fits entirely on a single computer screen so it can be monitored at a glance”.

In the present terms, a dashboard can be defined as a data visualization tool that displays the current status of metrics and key performance indicators (KPIs) simplifying complex data sets to provide users with at a glance awareness of current performance.

Dashboards consolidate and arrange numbers and metrics on a single screen. They can be tailored for a specific role and display metrics of a department or an organization on the whole.

Dashboards can be static for a one-time view, or dynamic showing the consolidated results of the data changes behind the screen. They can also be made interactive to display the various segments of large data on a single screen.

## Key Metrics for Dashboard

The core of the dashboard lies in the key metrics required for monitoring. Thus, based on whether the dashboard is for an organization on the whole or for a department such as sales, finance, human resources, production, etc. the key metrics that are required for display vary.

Further, the key metrics for a dashboard also depend on the role of the recipients (audience). For example, Executive (CEO, CIO, etc.), Operations Manager, Sales Head, Sales Manager, etc. This is due to the fact that the primary goal of a dashboard in to enable data visualization for decision making.

The success of a dashboard often depends on the metrics that were chosen for monitoring. For example, Key Performance Indicators, Balanced Scorecards and Sales Performance Figures could be the content appropriate in business dashboards.

## Dashboard Benefits

Dashboards allow managers to monitor the contribution of the various departments in the organization. To monitor the organization’s overall performance, dashboards allow you to capture and report specific data points from each of the departments in the organization, providing a snapshot of current performance and a comparison with earlier performance.

Benefits of dashboards include the following −

* Visual presentation of performance measures.
* Ability to identify and correct negative trends.
* Measurement of efficiencies/inefficiencies.
* Ability to generate detailed reports showing new trends.
* Ability to make more informed decisions based on collected data.
* Alignment of strategies and organizational goals.
* Instant visibility of all systems in total.
* Quick identification of data outliers and correlations.
* Time saving with the comprehensive data visualization as compared to running multiple reports.

## Types of Dashboards

Dashboards can be categorized based on their utility as follows −

* Strategic Dashboards
* Analytical Dashboards
* Operational Dashboards
* Informational Dashboards

### **Strategic Dashboards**

Strategic dashboards support managers at any level in an organization for decision making. They provide the snapshot of data, displaying the health and opportunities of the business, focusing on the high level measures of performance and forecasts.

* Strategic dashboards require to have periodic and static snapshots of data (e.g. daily, weekly, monthly, quarterly and annually). They need not be constantly changing from one moment to the next and require an update at the specified intervals of time.
* They portray only the high level data not necessarily giving the details.
* They can be interactive to facilitate comparisons and different views in case of large data sets at the click of a button. But, it is not necessary to provide more interactive features in these dashboards.

The following screenshot shows an example of an executive dashboard, displaying goals and progress.

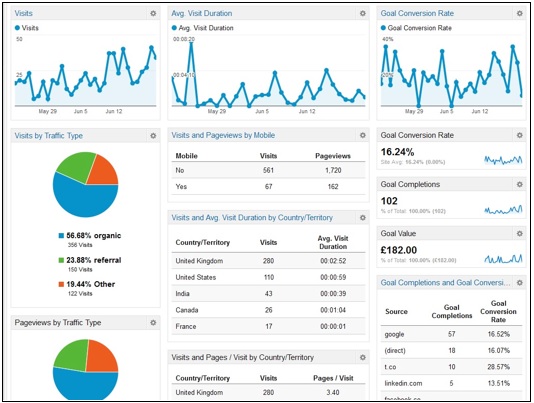


### **Analytical Dashboards**

Analytical dashboards include more context, comparisons, and history. They focus on the various facets of data required for analysis.

Analytical dashboards typically support interactions with the data, such as drilling down into the underlying details and hence should be interactive.

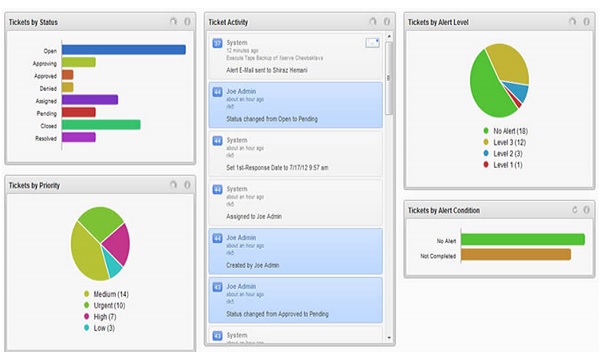
Examples of analytical dashboards include Finance Management dashboard and Sales Management dashboard.



### **Operational Dashboards**

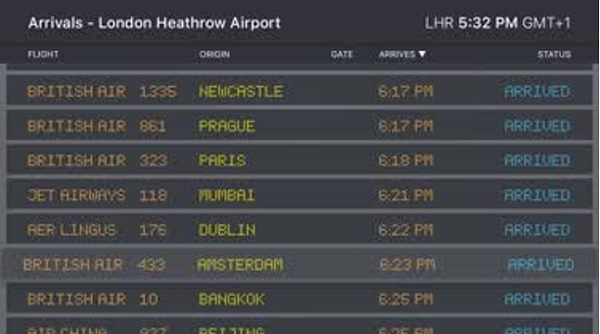
Operational dashboards are for constant monitoring of operations. They are often designed differently from strategic or analytical dashboards and focus on monitoring of activities and events that are constantly changing and might require attention and response at a moment's notice. Thus, operational dashboards require live and up to date data available at all times and hence should be dynamic.

An example of an operation dashboard could be a **support-system dashboard**, displaying live data on service tickets that require an immediate action from the supervisor on high-priority tickets.



**Informational Dashboards**

Informational dashboards are just for displaying figures, facts and/or statistics. They can be either static or dynamic with live data but not interactive. For example, flights arrival/departure information dashboard in an airport.

****

**Dashboard Data and Formats**

The data required for a dashboard depends on its category. The premise for the data is that it should be relevant, error-free, up to date and live if required. The data can possibly be from various and different sources and formats (Spreadsheets, Text Files, Web Pages, Organizational Database, etc.).

The results displayed on a dashboard must be authentic, correct and apt. This is crucial since the information on a dashboard would lead to decisions, actions and/or inferences. Thus, along with the data being displayed, the medium chosen for the display is equally important as it should not give an erroneous impression in the data portrayal. The focus should be on the ability of the data visualization that would unambiguously project the conclusions.

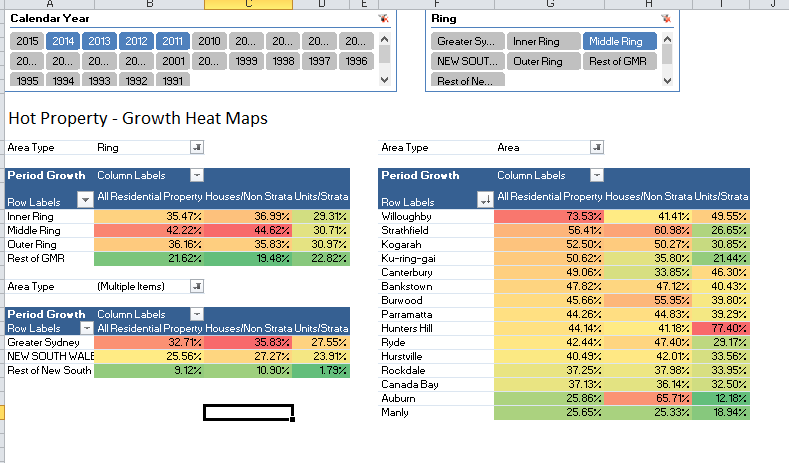
**Live Data on Dashboards**

As discussed earlier in this chapter, data warehousing and online analytical processing (OLAP) is making it possible to refresh the dynamic dashboards instantly with live data. It is also making those who design the dashboards be independent of the organization’s IT department for obtaining data.

Thus, the dashboards have become the most sought after medium from top management to a regular user.

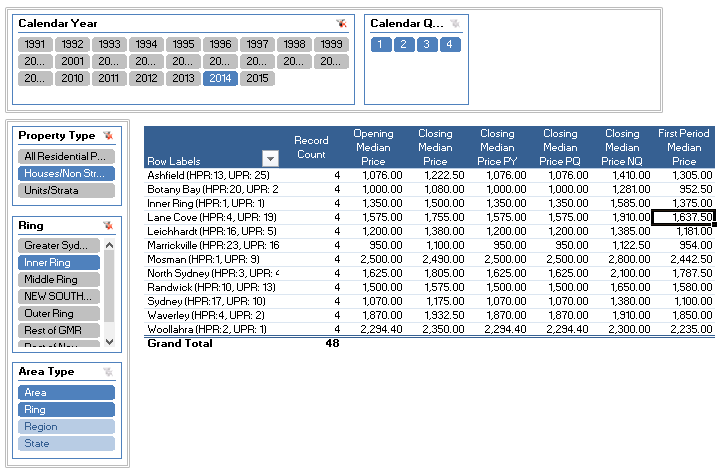
### **Growth sheet**

The Growth sheet uses Excel conditional formatting to provide a time effect for visualising growth for Solar and lunar eclipse. The user can adjust the time period or see more time in the table on the right of the sheet by interacting with the Slicers.



**Time Intelligence sheet**

Time Intelligence is a key feature of Power Pivot. Time Intelligence functions are used in many of the calculations in this model. This Time Intelligence sheet makes it easier for users to see the behaviour of different combinations of Time Intelligence functions against a single value.

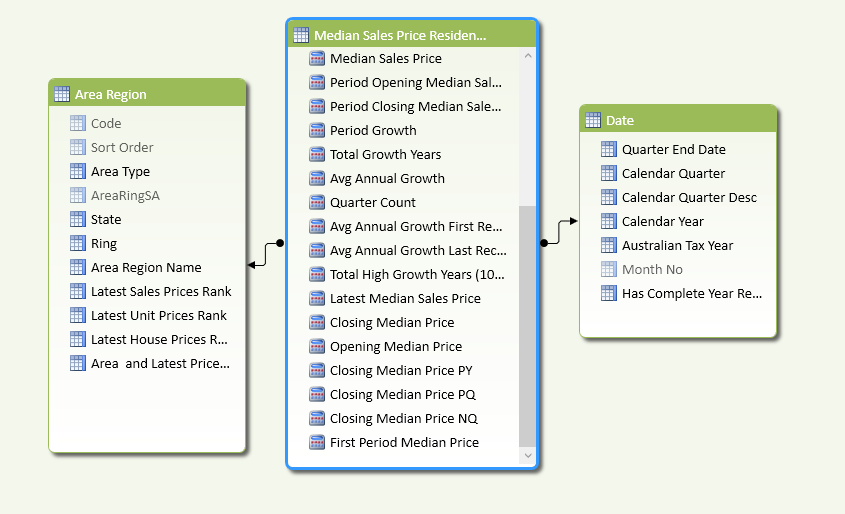


**Power Pivot model data**

The model uses data from the Solar and lunar eclipse . You can obtain the original data [here](http://www.housing.nsw.gov.au/About+Us/Reports+Plans+and+Papers/Rent+and+Sales+Reports/Latest+Issue/). I have manipulated this raw data using Power Query to create the final tables seen in the model. These tables are:

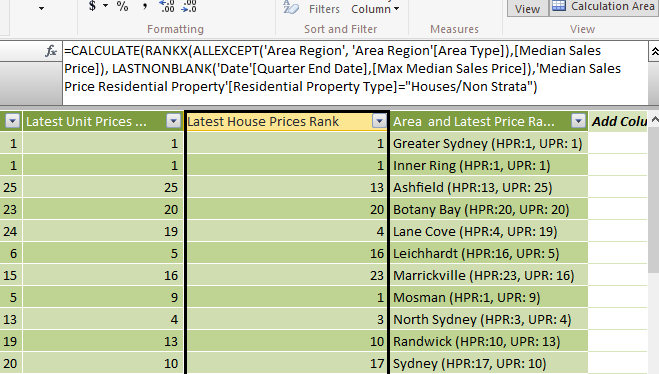
* **Area Region - which is a list of geographic regions at various levels within datasheet**
* **Median time- this table contains all the median latitude data for each quarter in the period and for each Area Region**
* **Date - a distinct list of quarter end dates used for the median time data**

The median price data in the model is non-aggregatable. It is not valid to add two median values together. Instead the calculations focus on comparing these median values over time. Although we cannot aggregate the median prices, the Area Region table contains geographic regions at various levels and the median prices for these aggregate region-levels are contained in the data set as pre-calculated values.



**Power Pivot model calculations**

The model defines a number of calculations using DAX. DAX is similar to Excel formulas but more powerful for advanced modelling scenarios. You can inspect the formula behind any of the values in the Pivot Tables by opening the Power Pivot window and selecting the Median sheet in the Data View from the Home tab. The bottom panel of this screen shows all the calculations and clicking on one will show the DAX formula in the formula bar

**  
Power Pivot window showing a DAX calculation in the formula bar**

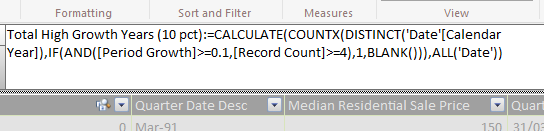
DAX formulas are very powerful and can be used to solve most modelling and analytics requirements. Some features include

Time Intelligence which can be used to calculated YTD, parallel periods, growth, and opening and closing balances

CALCULATE statement that can manipulate the context of a calculation for complex set based analysis or advanced comparisons and correlations

Text manipulation to add groupings, concatenate columns, or create additional descriptive attributes

Most Excel formulas can be translated into DAX formulas. DAX also has capability beyond standard Excel formulas.

**  
Power Pivot formula bar showing DAX calculation using the CALCULATE statement**

**Source of Dataset:-**

**Link of dataset:-**

<https://www.kaggle.com/nasa/solar-eclipses>

# **Context**

Eclipses of the sun can only occur when the moon is near one of its two orbital nodes during the new moon phase. It is then possible for the Moon's penumbral, umbral, or antumbral shadows to sweep across Earth's surface thereby producing an eclipse. There are four types of solar eclipses: a partial eclipse, during which the moon's penumbral shadow traverses Earth and umbral and antumbral shadows completely miss Earth; an annular eclipse, during which the moon's antumbral shadow traverses Earth but does not completely cover the sun; a total eclipse, during which the moon's umbral shadow traverses Earth and completely covers the sun; and a hybrid eclipse, during which the moon's umbral and antumbral shadows traverse Earth and annular and total eclipses are visible in different locations. Earth will experience 11898 solar eclipses during the five millennium period -1999 to +3000 (2000 BCE to 3000 CE).

Eclipses of the moon can occur when the moon is near one of its two orbital nodes during the full moon phase. It is then possible for the moon to pass through Earth's penumbral or umbral shadows thereby producing an eclipse. There are three types of lunar eclipses: a penumbral eclipse, during which the moon traverses Earth's penumbral shadow but misses its umbral shadow; a partial eclipse, during which the moon traverses Earth's penumbral and umbral shadows; and a total eclipse, during which the moon traverses Earth's penumbral and umbral shadows and passes completely into Earth's umbra. Earth will experience 12064 lunar eclipses during the five millennium period -1999 to +3000 (2000 BCE to 3000 CE).

**ETL PROCESS:-**

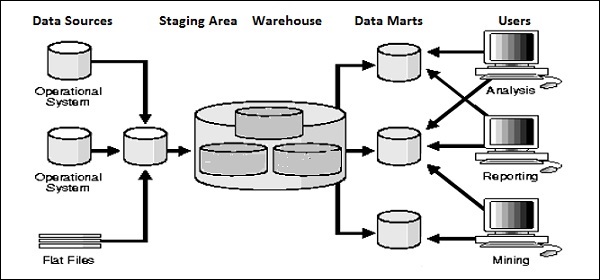
ETL stands for Extract, Transform and Load. An ETL tool extracts the data from different RDBMS source systems, transforms the data like applying calculations, concatenate, etc. and then load the data to Data Warehouse system. The data is loaded in the DW system in the form of dimension and fact tables.

**Extraction**

A staging area is required during ETL load. There are various reasons why staging area is required.

The source systems are only available for specific period of time to extract data. This period of time is less than the total data-load time. Therefore, staging area allows you to extract the data from the source system and keeps it in the staging area before the time slot ends.

* Staging area is required when you want to get the data from multiple data sources together or if you want to join two or more systems together. For example, you will not be able to perform a SQL query joining two tables from two physically different databases.
* Data extractions’ time slot for different systems vary as per the time zone and operational hours.
* Data extracted from source systems can be used in multiple data warehouse system, Operation Data stores, etc.
* ETL allows you to perform complex transformations and requires extra area to store the data.



**Transform**

In data transformation, you apply a set of functions on extracted data to load it into the target system. Data, which does not require any transformation is known as direct move or pass through data.

You can apply different transformations on extracted data from the source system. For example, you can perform customized calculations. If you want sum-of-sales revenue and this is not in database, you can apply the SUM formula during transformation and load the data.

For example, if you have the first name and the last name in a table in different columns, you can use concatenate before loading.

**Load**

During Load phase, data is loaded into the end-target system and it can be a flat file or a Data Warehouse system.

**What is ETL?**

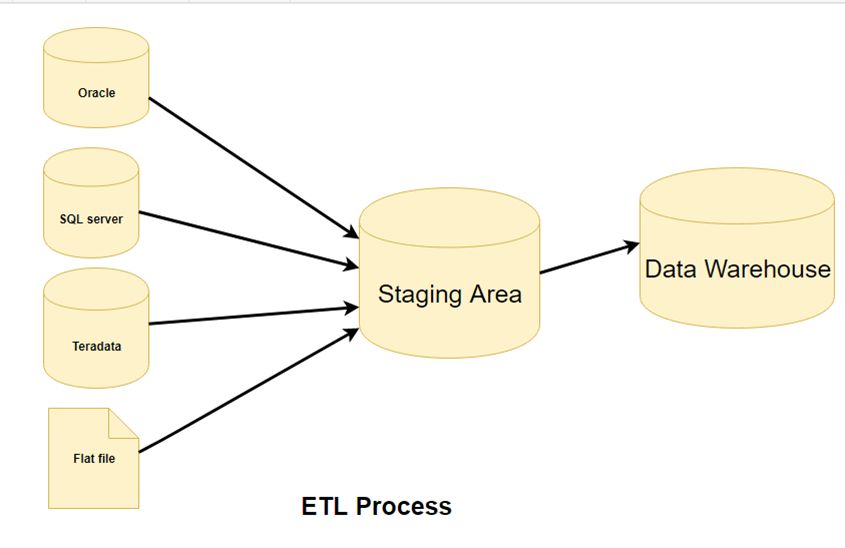
**ETL** is a process that extracts the data from different source systems, then transforms the data (like applying calculations, concatenations, etc.) and finally loads the data into the Data Warehouse system. Full form of ETL is Extract, Transform and Load.

It's tempting to think a creating a Data warehouse is simply extracting data from multiple sources and loading into database of a Data warehouse. This is far from the truth and requires a complex ETL process. The ETL process requires active inputs from various stakeholders including developers, analysts, testers, top executives and is technically challenging.

In order to maintain its value as a tool for decision-makers, Data warehouse system needs to change with business changes. ETL is a recurring activity (daily, weekly, monthly) of a Data warehouse system and needs to be agile, automated, and well documented.

## ETL Process in Data Warehouses

ETL is a 3-step process

[](https://www.guru99.com/images/1/022218_0848_ETLExtractT1.png)

## Step 1) Extraction

In this step, data is extracted from the source system into the staging area. Transformations if any are done in staging area so that performance of source system in not degraded. Also, if corrupted data is copied directly from the source into Data warehouse database, rollback will be a challenge. Staging area gives an opportunity to validate extracted data before it moves into the Data warehouse.

Data warehouse needs to integrate systems that have different

DBMS, Hardware, Operating Systems and Communication Protocols. Sources could include legacy applications like Mainframes, customized applications, Point of contact devices like ATM, Call switches, text files, spreadsheets, ERP, data from vendors, partners amongst others.

Hence one needs a logical data map before data is extracted and loaded physically. This data map describes the relationship between sources and target data.

**Three Data Extraction methods:**

1. Full Extraction
2. Partial Extraction- without update notification.
3. Partial Extraction- with update notification

Irrespective of the method used, extraction should not affect performance and response time of the source systems. These source systems are live production databases. Any slow down or locking could effect company's bottom line.

**Some validations are done during Extraction:**

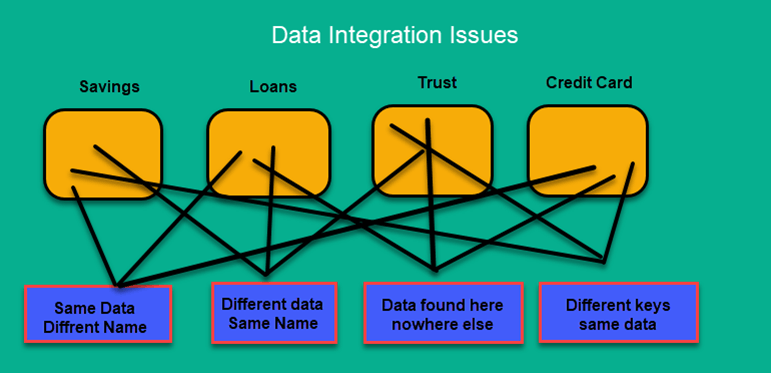
* Reconcile records with the source data
* Make sure that no spam/unwanted data loaded
* Data type check
* Remove all types of duplicate/fragmented data
* Check whether all the keys are in place or not

## Step 2) Transformation

Data extracted from source server is raw and not usable in its original form. Therefore it needs to be cleansed, mapped and transformed. In fact, this is the key step where ETL process adds value and changes data such that insightful BI reports can be generated.

In this step, you apply a set of functions on extracted data. Data that does not require any transformation is called as **direct move** or **pass through data**.

In transformation step, you can perform customized operations on data. For instance, if the user wants sum-of-sales revenue which is not in the database. Or if the first name and the last name in a table is in different columns. It is possible to concatenate them before loading.

[](https://www.guru99.com/images/1/022218_0848_ETLExtractT2.png)

**Following are Data Integrity Problems:**

1. Different spelling of the same person like Jon, John, etc.
2. There are multiple ways to denote company name like Google, Google Inc.
3. Use of different names like Cleaveland, Cleveland.
4. There may be a case that different account numbers are generated by various applications for the same customer.
5. In some data required files remains blank
6. Invalid product collected at POS as manual entry can lead to mistakes.

**Validations are done during this stage**

* Filtering – Select only certain columns to load
* Using rules and lookup tables for Data standardization
* Character Set Conversion and encoding handling
* Conversion of Units of Measurements like Date Time Conversion, currency conversions, numerical conversions, etc.
* Data threshold validation check. For example, age cannot be more than two digits.
* Data flow validation from the staging area to the intermediate tables.
* Required fields should not be left blank.
* Cleaning ( for example, mapping NULL to 0 or Gender Male to "M" and Female to "F" etc.)
* Split a column into multiples and merging multiple columns into a single column.
* Transposing rows and columns,
* Use lookups to merge data
* Using any complex data validation (e.g., if the first two columns in a row are empty then it automatically reject the row from processing)

## Step 3) Loading

Loading data into the target datawarehouse database is the last step of the ETL process. In a typical Data warehouse, huge volume of data needs to be loaded in a relatively short period (nights). Hence, load process should be optimized for performance.

In case of load failure, recover mechanisms should be configured to restart from the point of failure without data integrity loss. Data Warehouse admins need to monitor, resume, cancel loads as per prevailing server performance.

**Types of Loading:**

* **Initial Load** — populating all the Data Warehouse tables
* **Incremental Load**— applying ongoing changes as when needed periodically.
* **Full Refresh** —erasing the contents of one or more tables and reloading with fresh data.

### Load verification

* Ensure that the key field data is neither missing nor null.
* Test modeling views based on the target tables.
* Check that combined values and calculated measures.
* Data checks in dimension table as well as history table.
* Check the BI reports on the loaded fact and dimension table.

## ETL tools

There are many Data Warehousing tools are available in the market. Here, are some most prominent one.

**1. MarkLogic:**

MarkLogic is a data warehousing solution which makes data integration easier and faster using an array of enterprise features. It can query different types of data like documents, relationships, and metadata.

<https://developer.marklogic.com/products/>

**2. Oracle:**

Oracle is the industry-leading database. It offers a wide range of choice of Data Warehouse solutions for both on-premises and in the cloud. It helps to optimize customer experiences by increasing operational efficiency.

<https://www.oracle.com/index.html>

**3. Amazon RedShift:**

Amazon Redshift is Datawarehouse tool. It is a simple and cost-effective tool to analyze all types of data using standard SQL and existing BI tools. It also allows running complex queries against petabytes of structured data.

<https://aws.amazon.com/redshift/?nc2=h_m1>

Here is a complete list of useful [Data warehouse Tools.](https://www.guru99.com/top-20-etl-database-warehousing-tools.html)

## Best practices ETL process

**Never try to cleanse all the data:**

Every organization would like to have all the data clean, but most of them are not ready to pay to wait or not ready to wait. To clean it all would simply take too long, so it is better not to try to cleanse all the data.

**Never cleanse Anything:**

Always plan to clean something because the biggest reason for building the Data Warehouse is to offer cleaner and more reliable data.

**Determine the cost of cleansing the data:**

Before cleansing all the dirty data, it is important for you to determine the cleansing cost for every dirty data element.

**To speed up query processing, have auxiliary views and indexes:**

To reduce storage costs, store summarized data into disk tapes. Also, the trade-off between the volume of data to be stored and its detailed usage is required. Trade-off at the level of granularity of data to decrease the storage costs.

## Summary:

* ETLstands for Extract, Transform and Load.
* ETL provides a method of moving the data from various sources into a data warehouse.
* In the first step extraction, data is extracted from the source system into the staging area.
* In the transformation step, the data extracted from source is cleansed and transformed .
* Loading data into the target data warehouse is the last step of the ETL process

**Analysis on dataset (for each analysis):-**

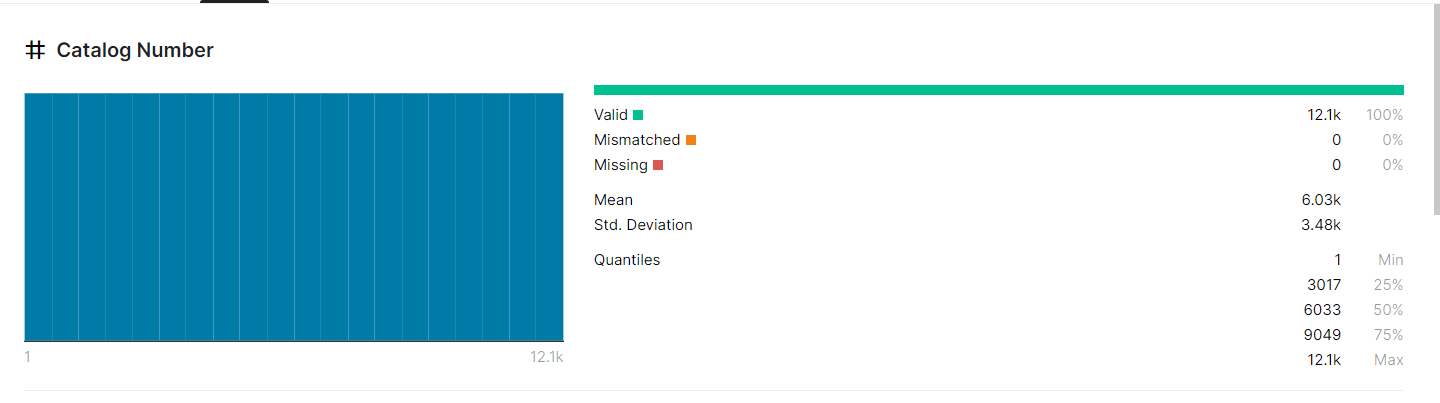
**1.INTRODUCTION**

The sheets which are given by teacher is all about the solar eclipse and lunar eclipse. In this sheet we could see many columns like calendar date which is sub divided using

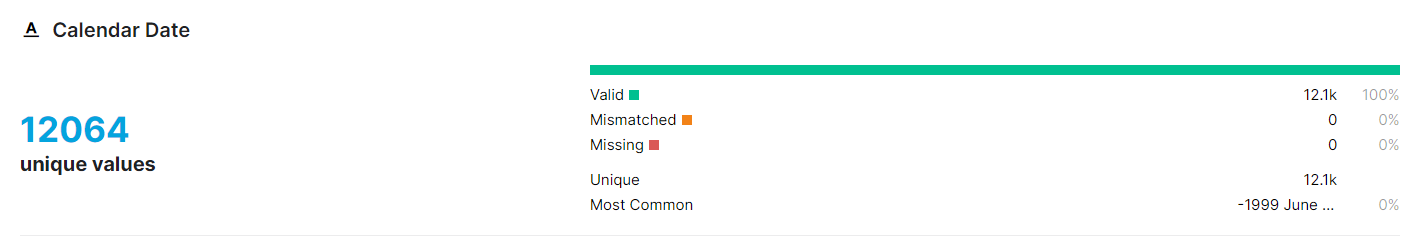
Split delimiter, eclipse which is a raw data is converted to a specified time, the delta value, lunation number, saros number, eclipse type, quincena solar eclipse gamma penumbral magnitude ,umbral magnitude, latitude, longitude, Penumbral Eclipse Duration (m), Partial Eclipse Duration (m), Total Eclipse Duration (m).

**Overview of sheet:-**

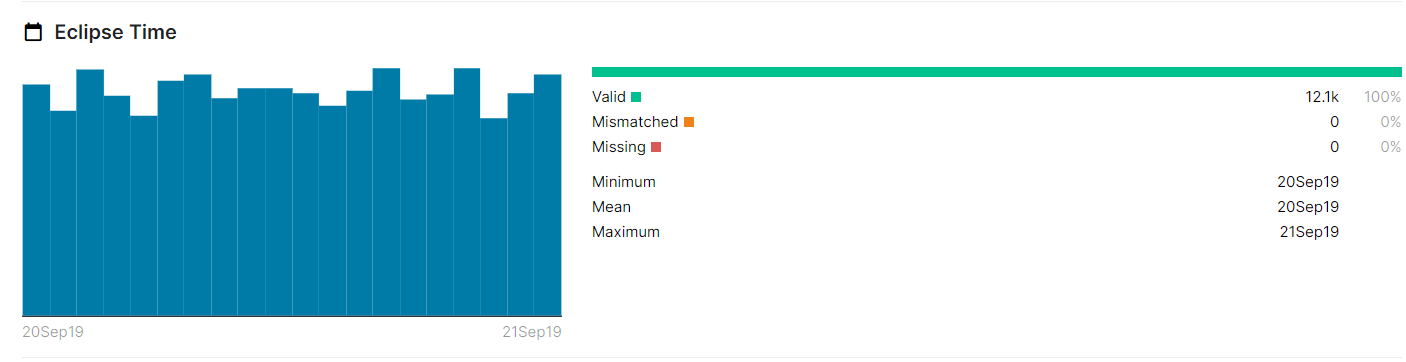
**1.Catalog Number**

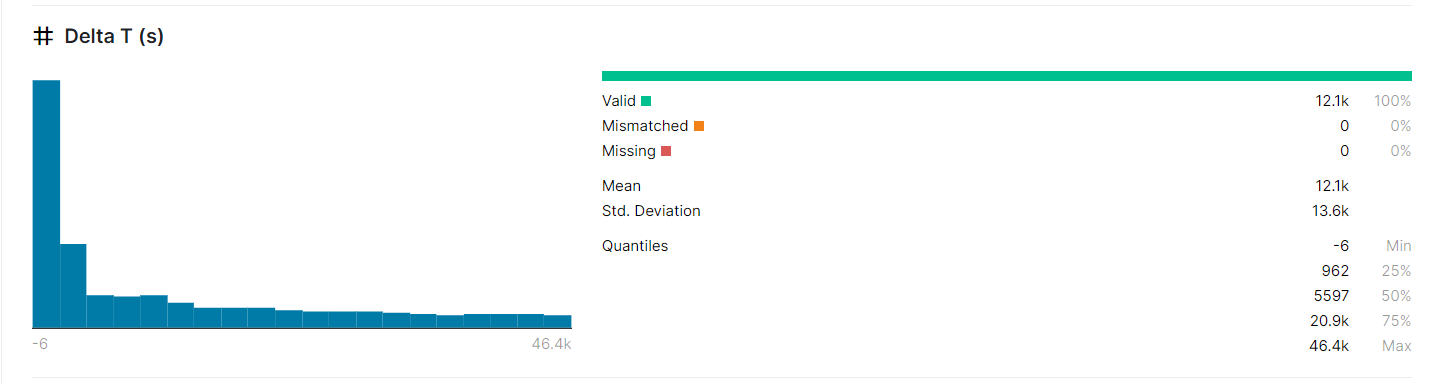


**2.Calendar Date**

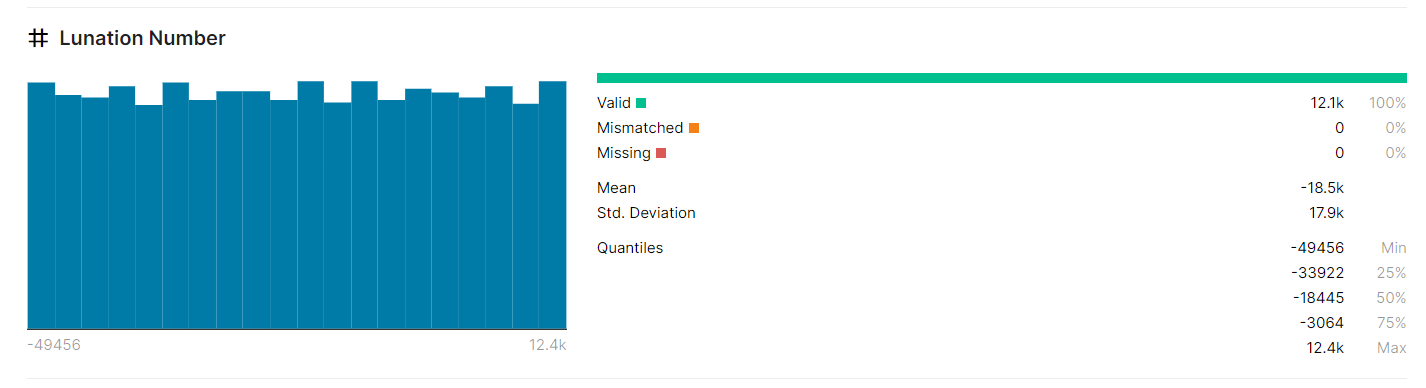


**3.Eclipse Time:-**

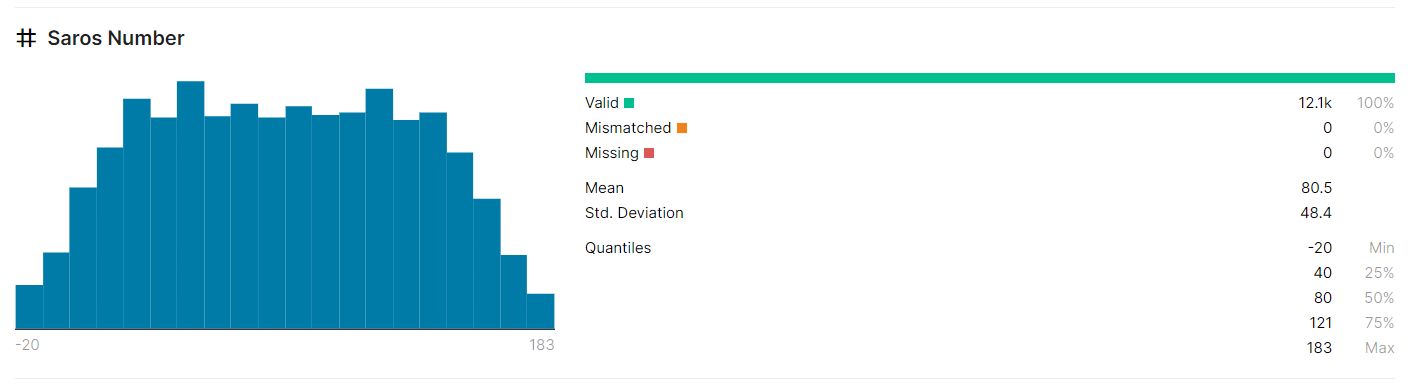
**4.Delta**



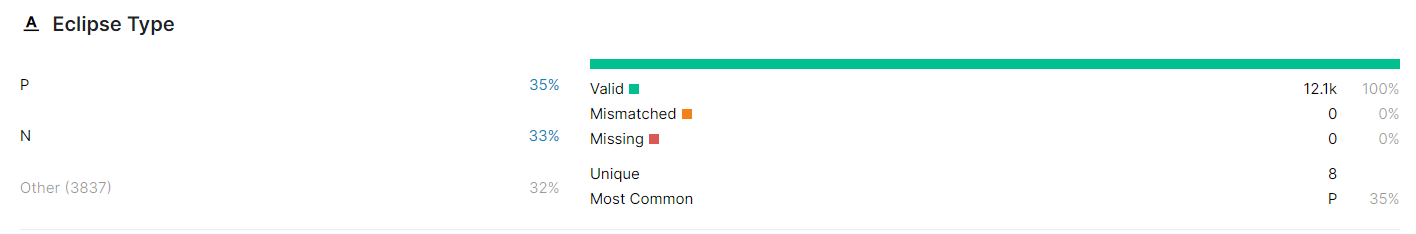
**5.Lunation Number**



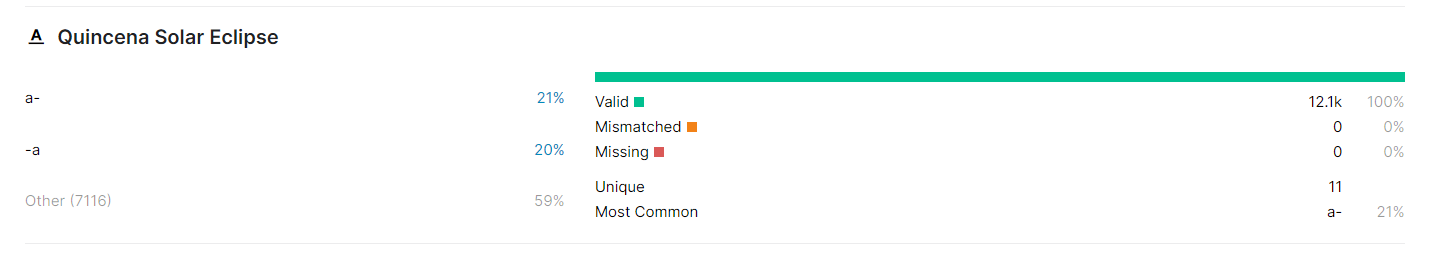
**6.Saros Number**



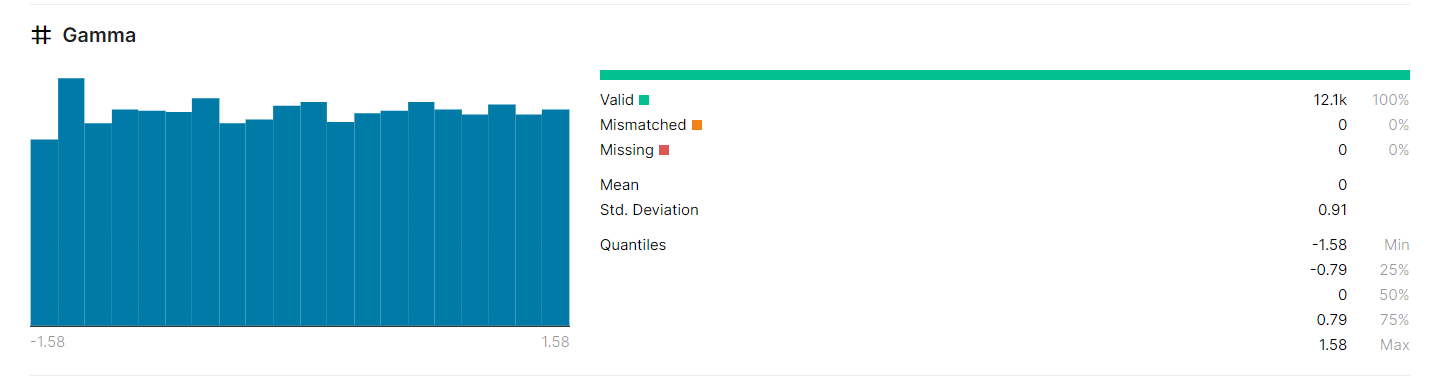
**7.Eclipse Type**



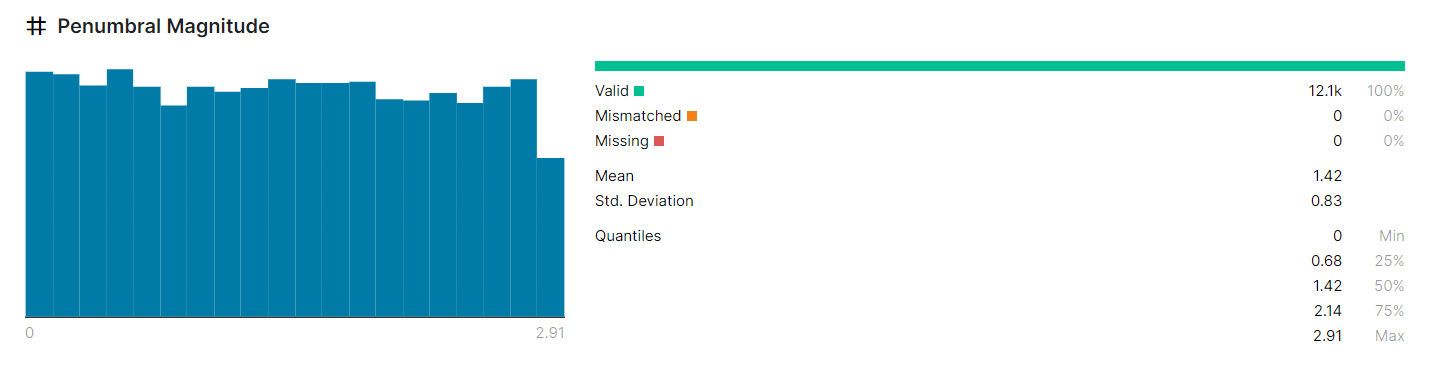
**8.Quincena Solar Eclipse**



**9.Gamma**



**10.Penumbral Magnitude**



**General Description:-**

Microsoft Excel is a helpful and powerful program for data analysis and documentation. It is a spreadsheet program, which contains a number of columns and rows, where each intersection of a column and a row is a “cell.” Each cell contains one point of data or one piece of information.

**Tableau** is a data visualization software that is used for data science and business intelligence. ... It comes with tools that allow to drill down data and see the impact in a visual format that can be easily understood by any individual. **Tableau** also comes with real-time data analytics capabilities and cloud support.

A **data set** (or **dataset**) is a collection of data. In the case of tabular data, a **data set** corresponds to one or more **database** tables, where every column of a table represents a particular variable, and each row corresponds to a given record of the **data set** in question.

In **Excel**, a **formula** is an expression that operates on values in a range of cells or a cell. For example, =A1+A2+A3, which finds the sum of the range of values from cell A1 to cell A3.

A **pivot table** is a statistics tool that summarizes and reorganizes selected columns and rows of data in a spreadsheet or database **table** to obtain a desired report. The tool does not actually change the spreadsheet or database itself, it simply “pivots” or turns the data to view it from different perspectives.

**Specific Requirements, functions and formulas**

1.Pivot tables

2.median

3.mean

4.count

5.sum

6.avg

7.split

8.pivot charts

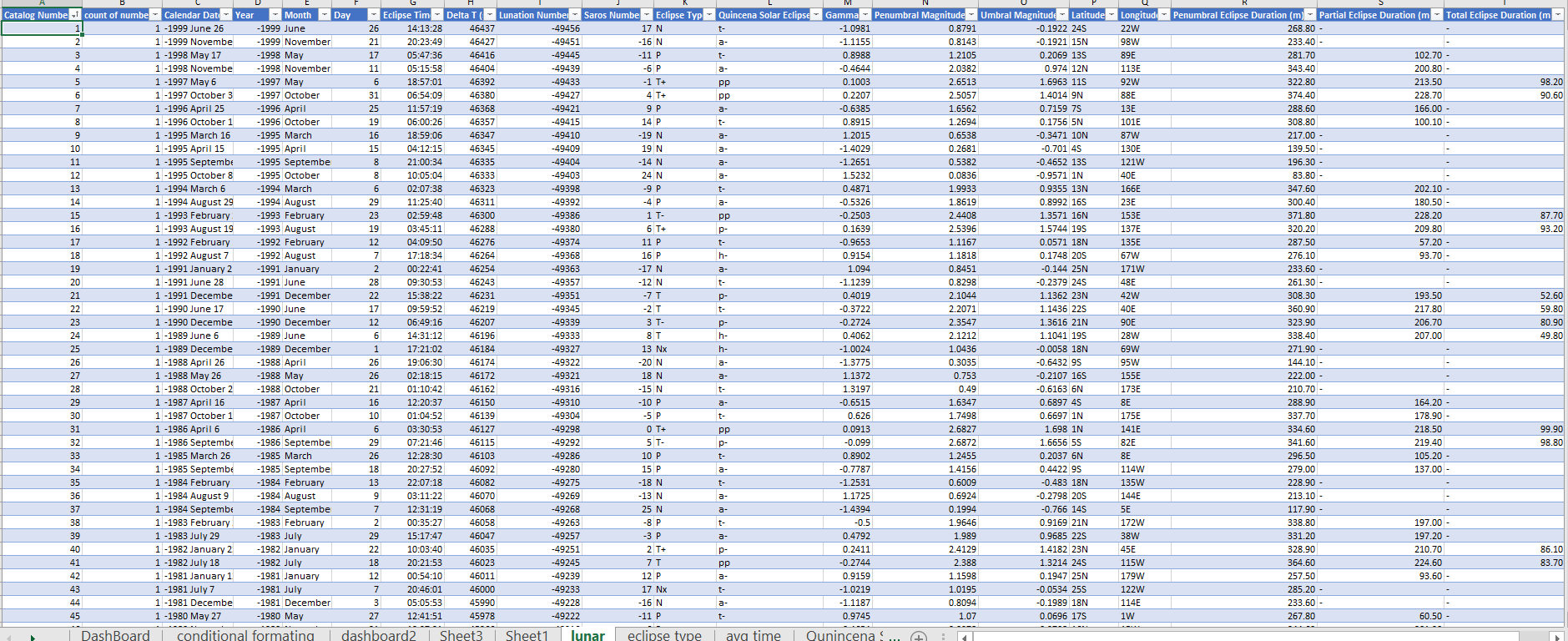
9.add

10.std.dev

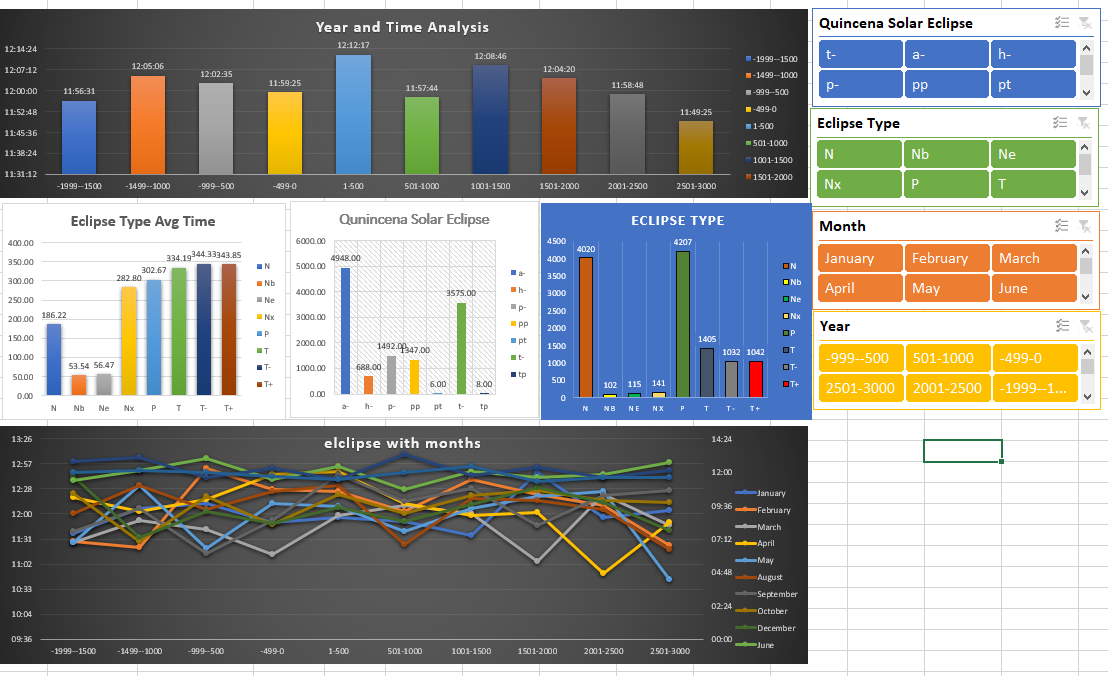
11.txt

12.etl process etc…

**Given Data sheet:-**

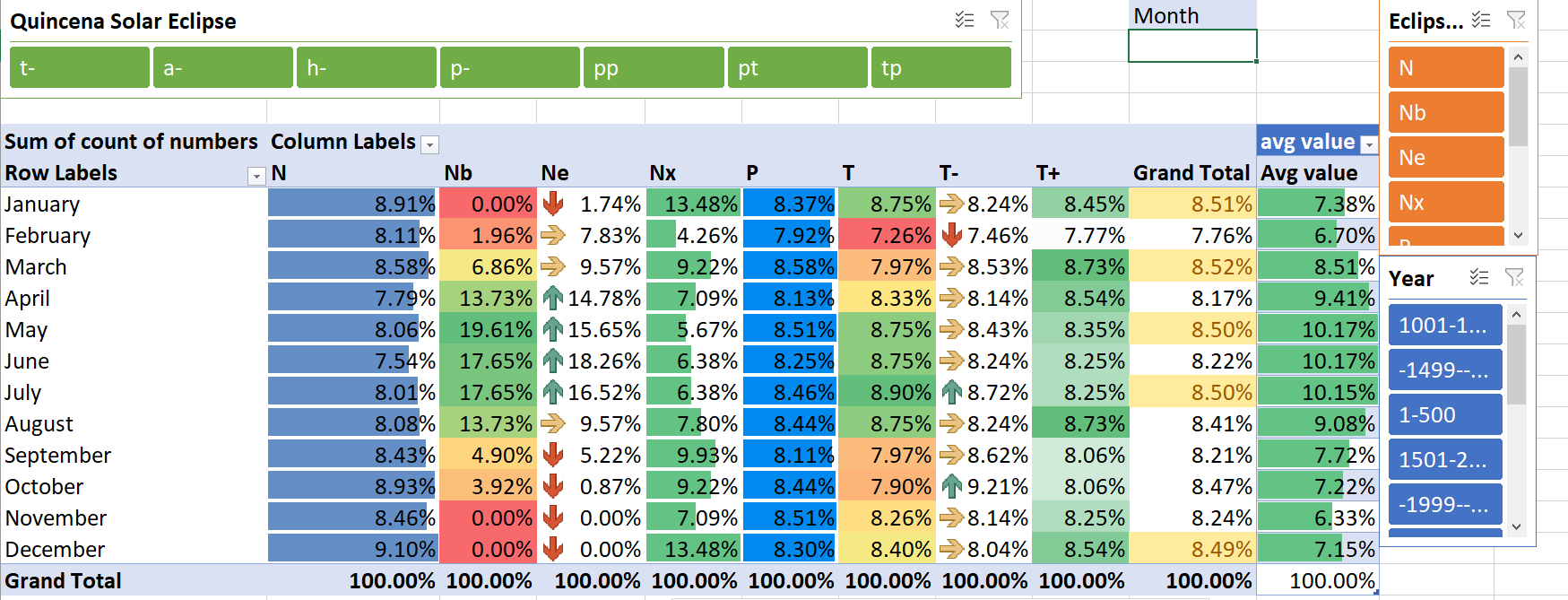


**Anayisis results:-**

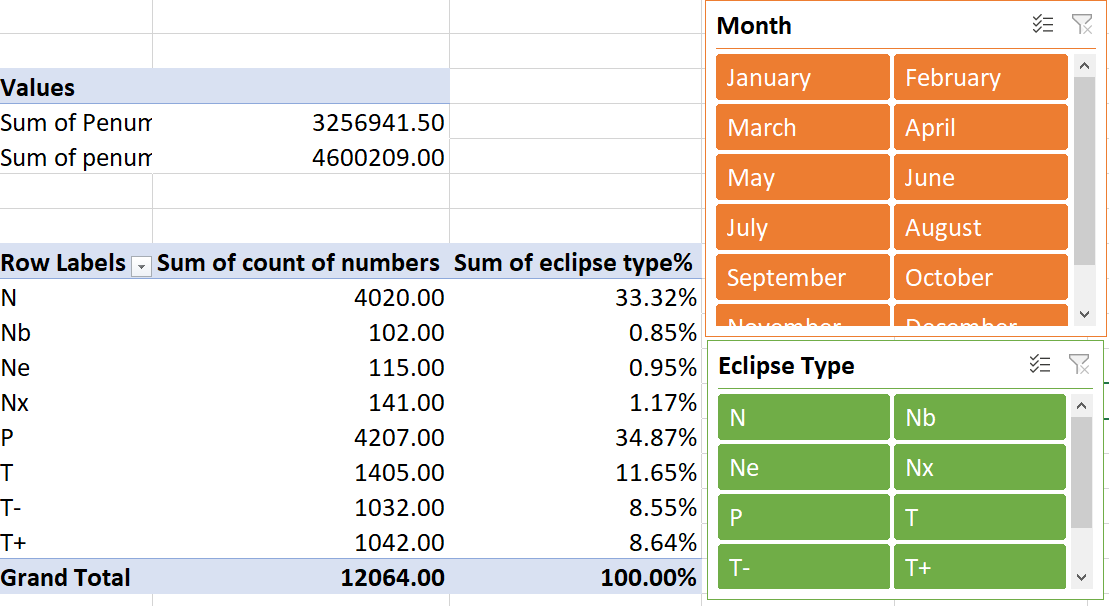


This is the basic mix of all ananlyis.

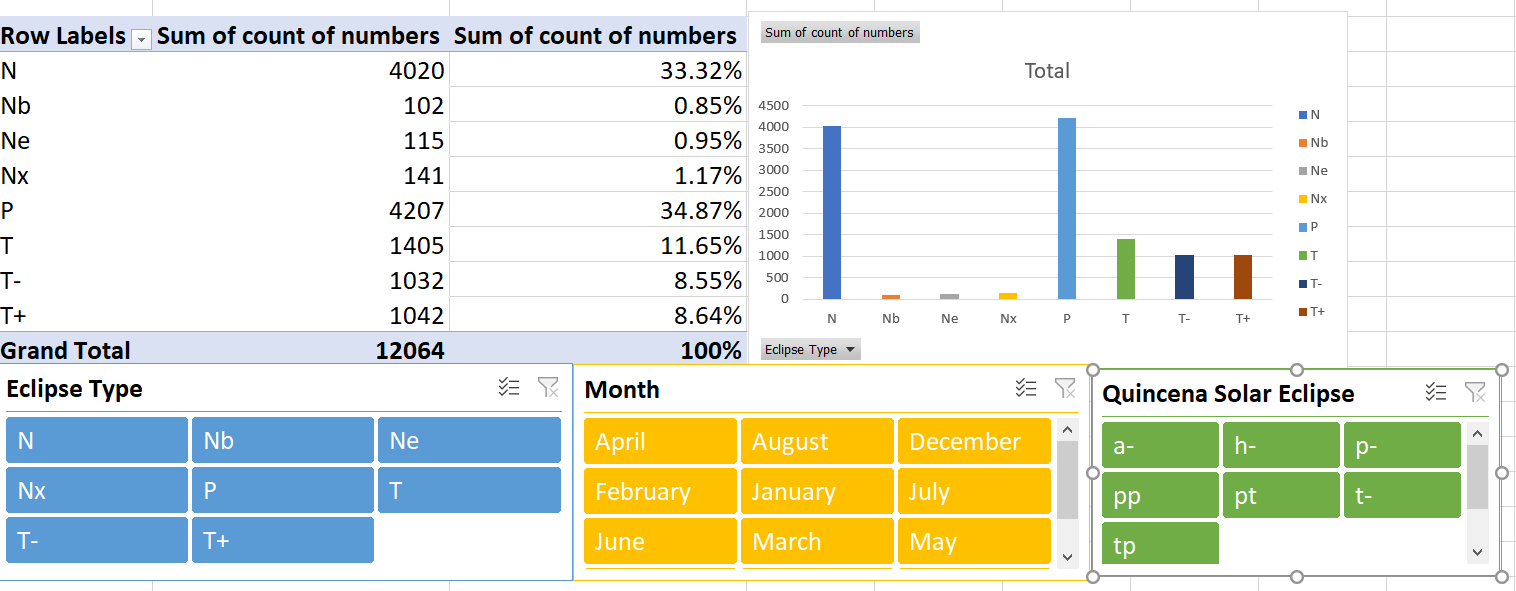
**The total conditional formating**



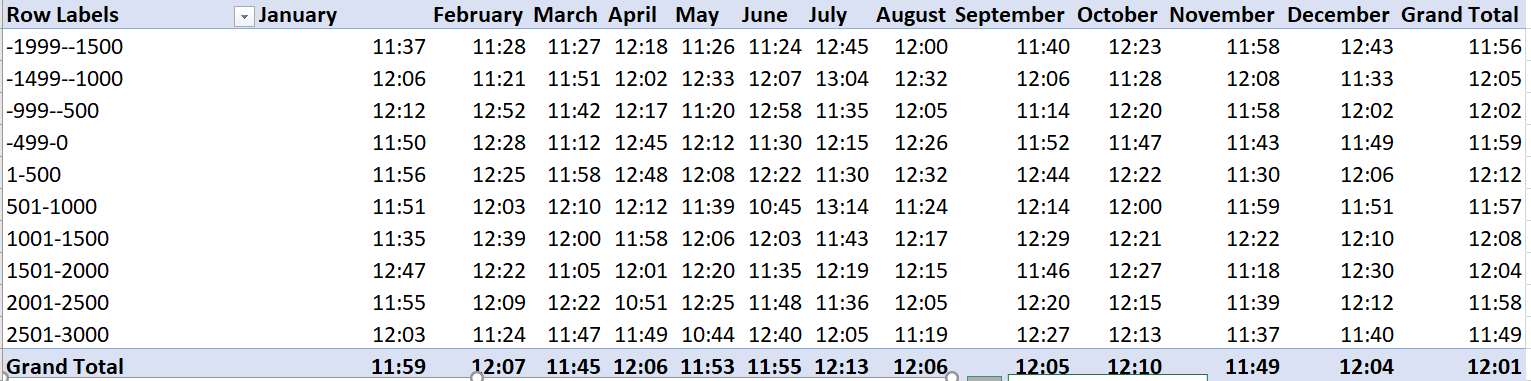
The calculation of eclipse type and number of similar types and avg of percentage

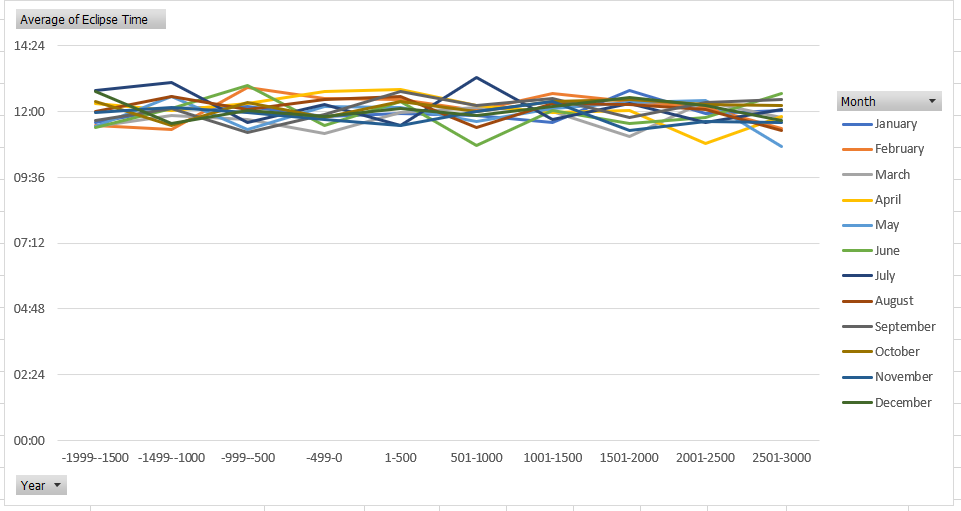


Same calculation in different chart

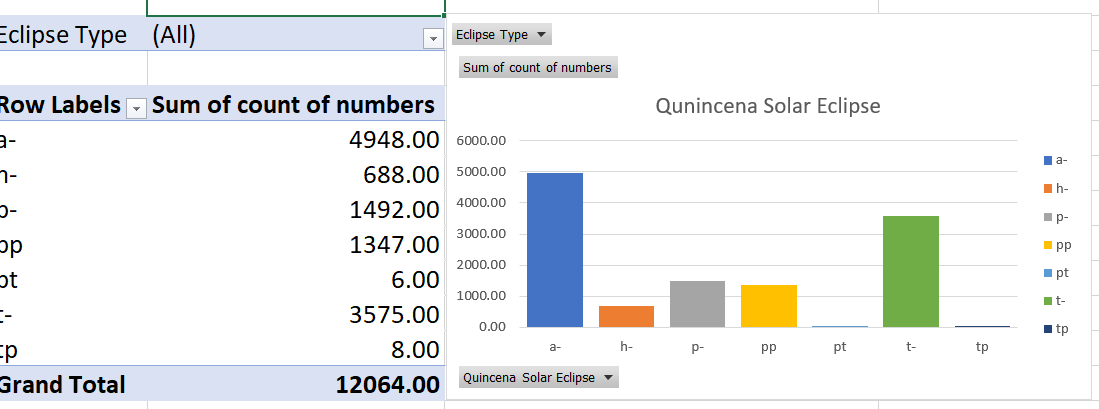


The calculation of grouped years and months with the values of mean time

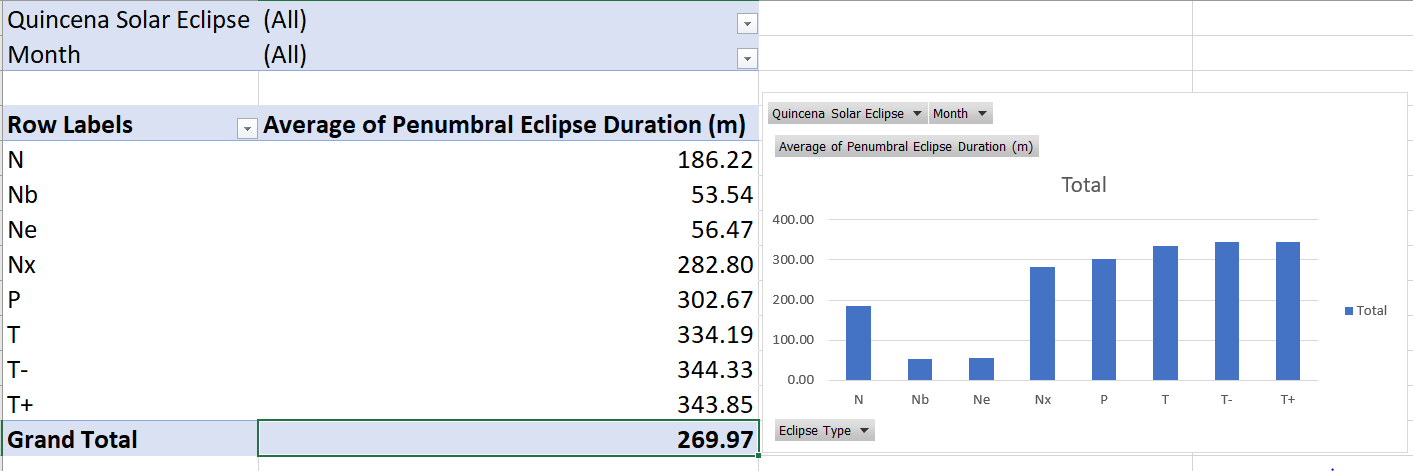




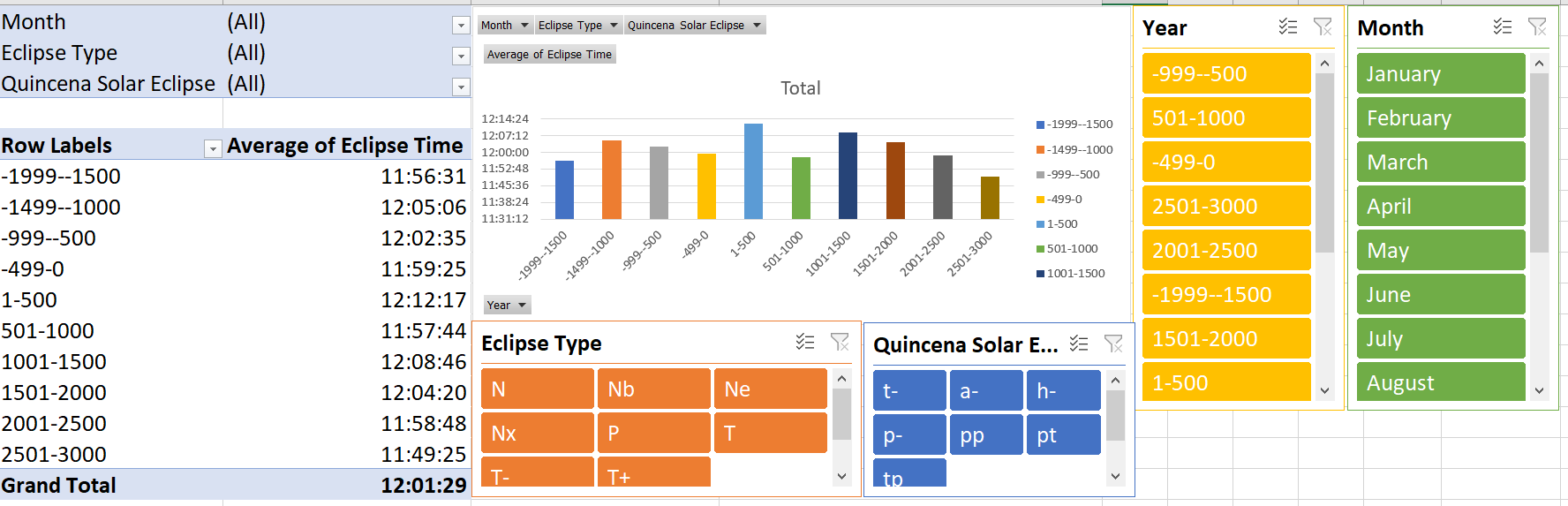
Quincena Solar eclipse type calculation:



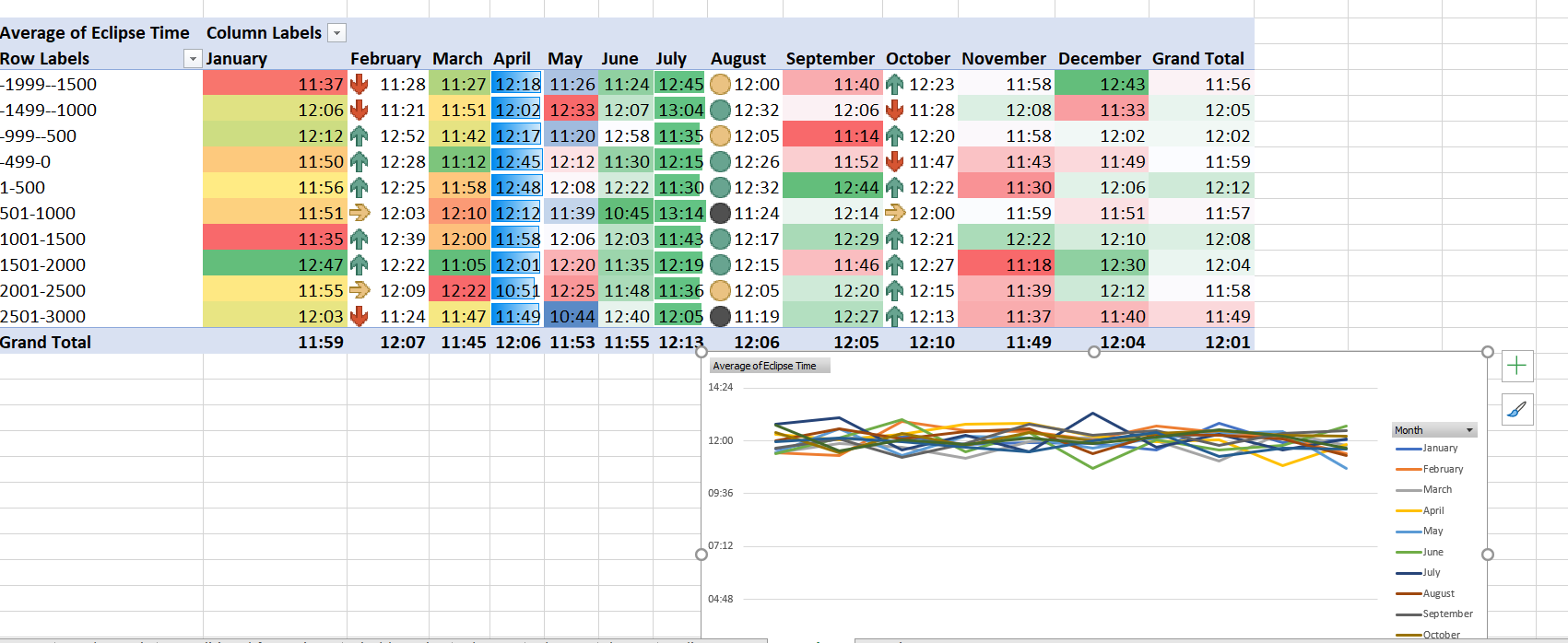
The calculation of eclipse type with avg penumbral eclipse duration



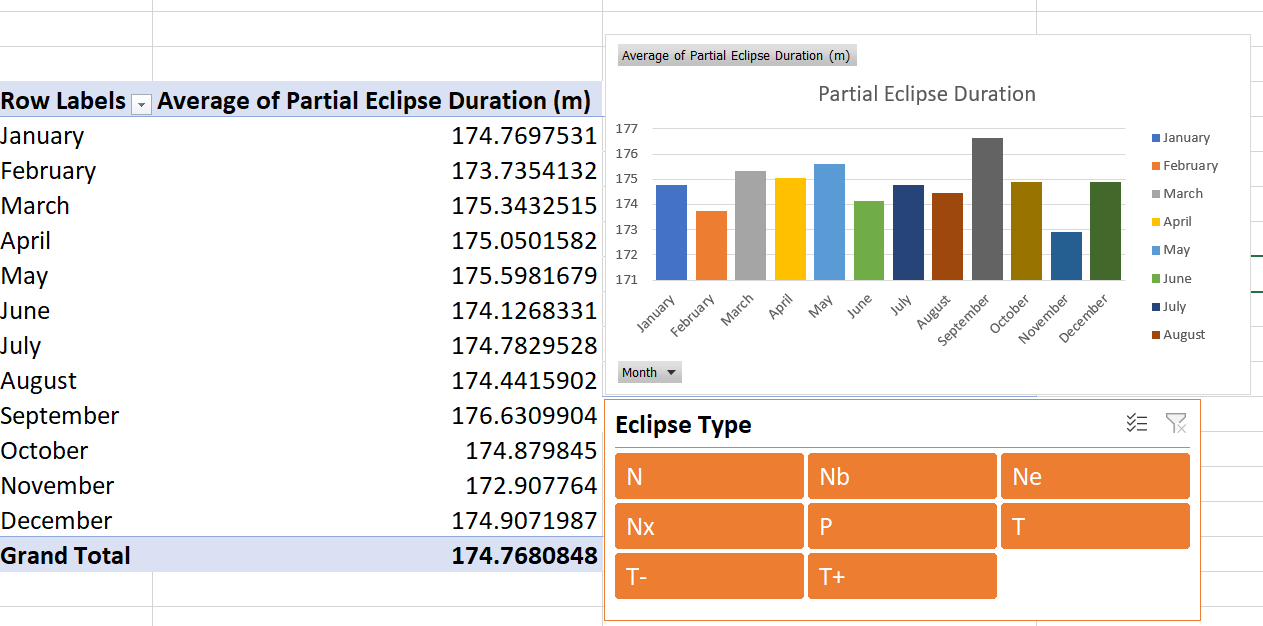
The calculation of eclipse grouped area with Avgerage eclipse type.



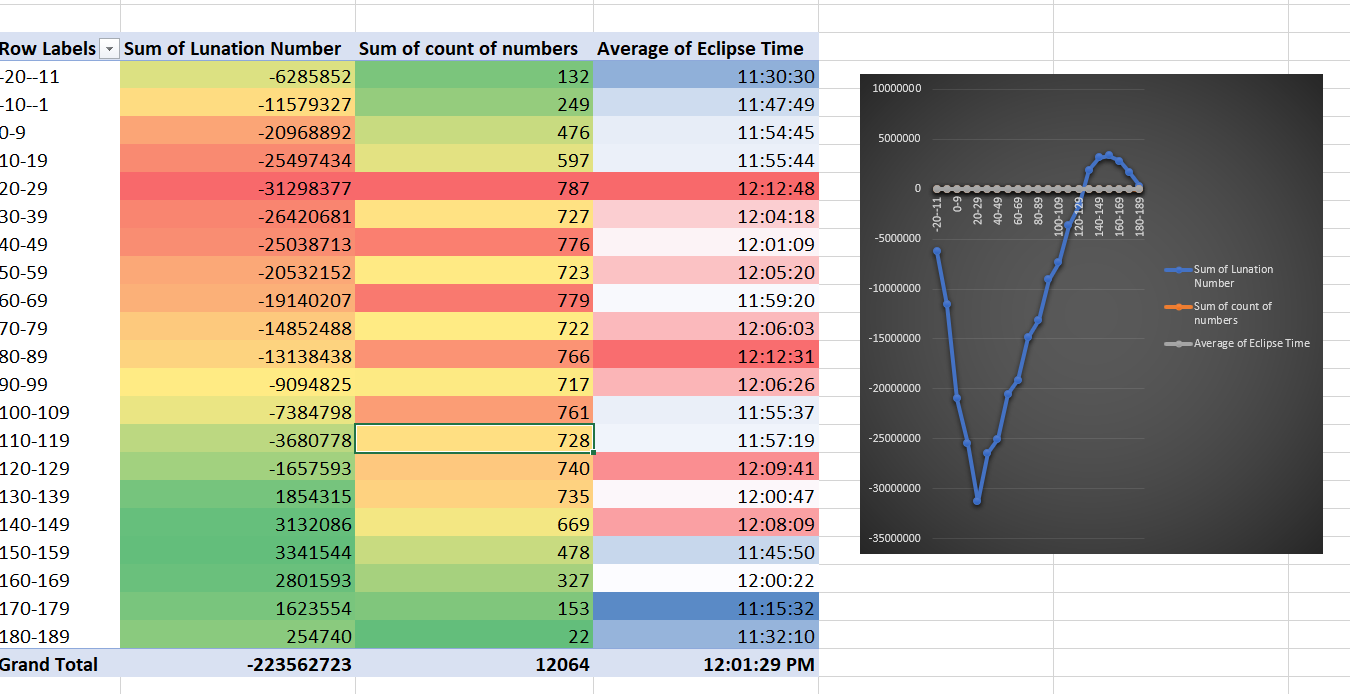
Avg of the eclipse time with respect to year and month



The calculation of eclipse month with Avgerage Partial eclipse type.



Analysis of year lunation number and avg eclipse time.



Total analysis of lunar eclipse.

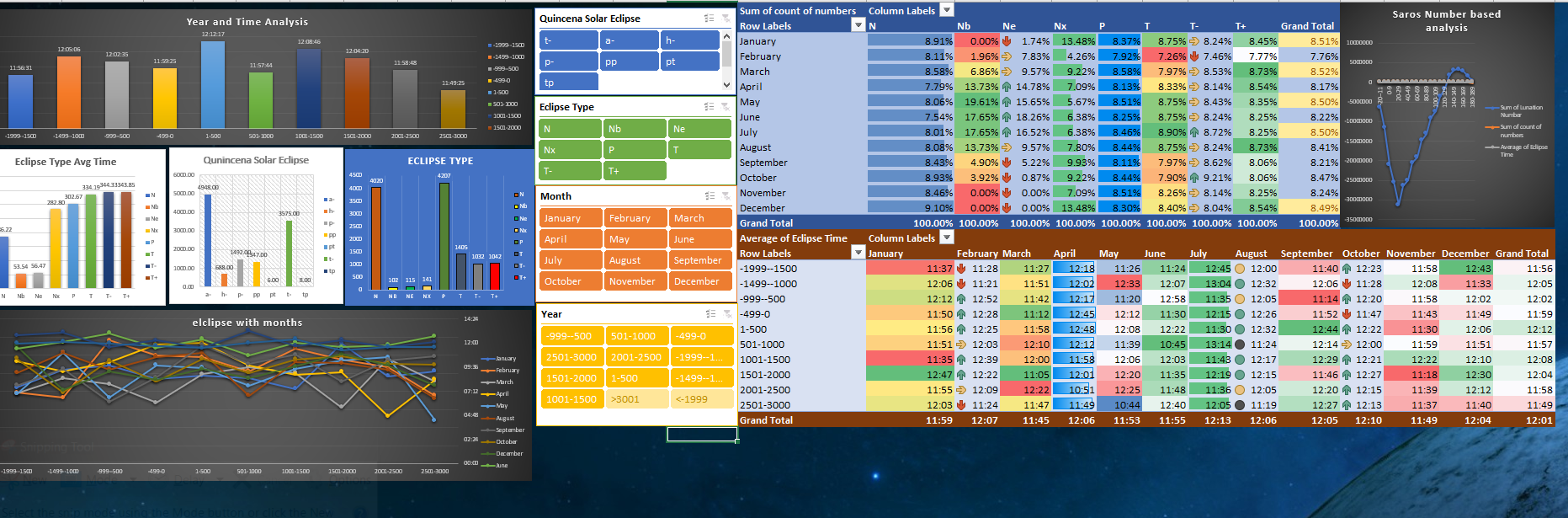
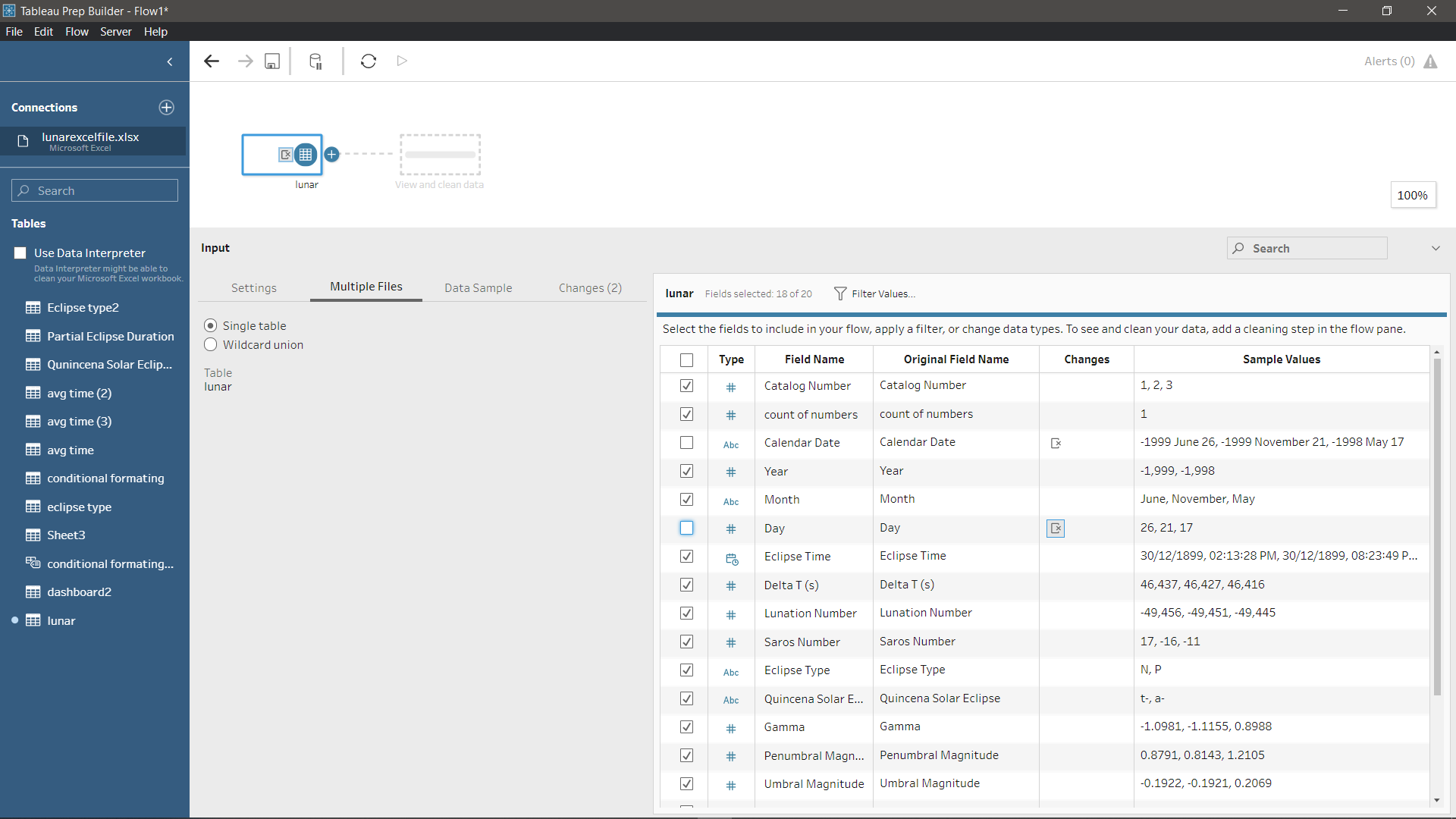
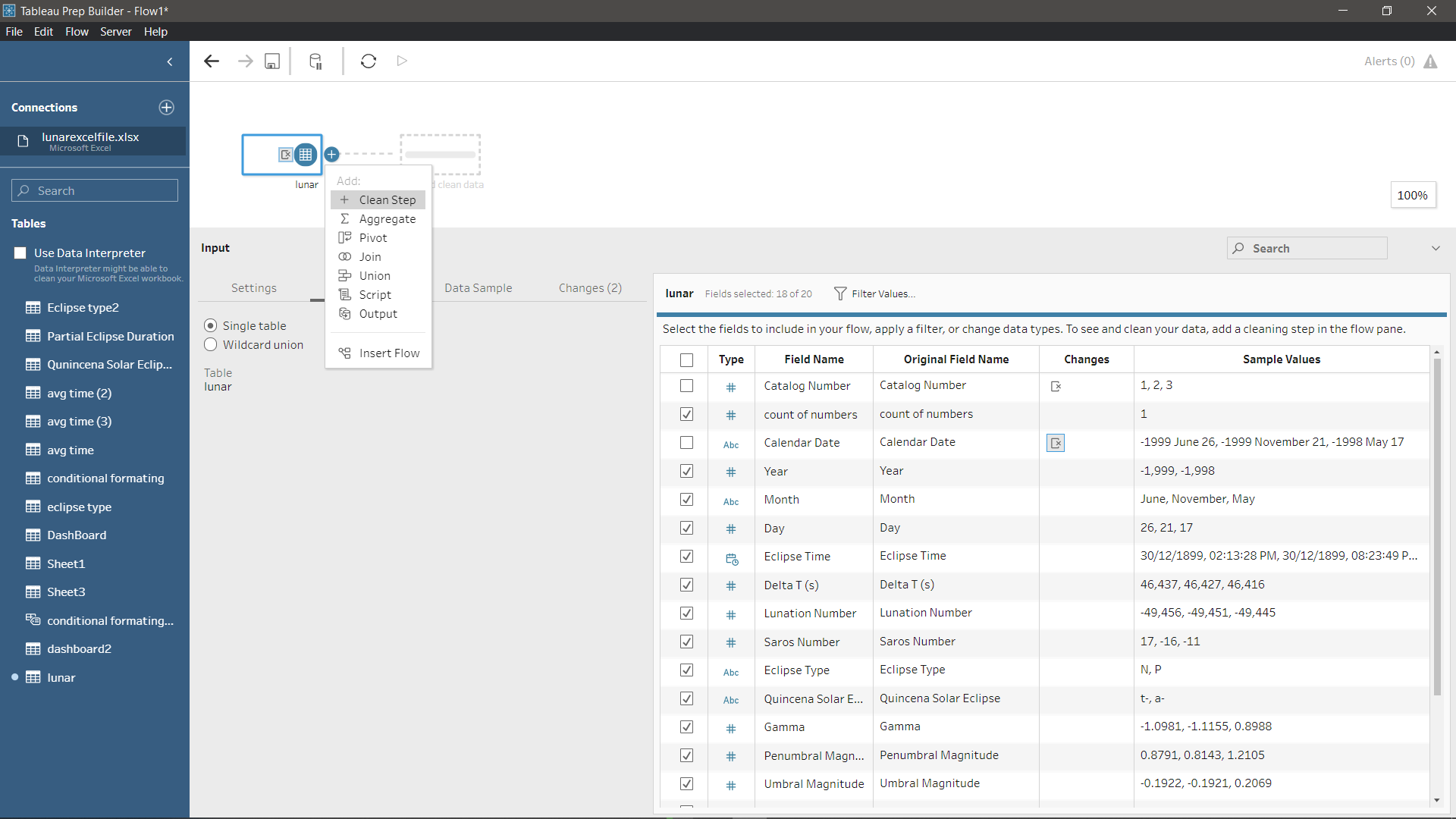


Tableau:-

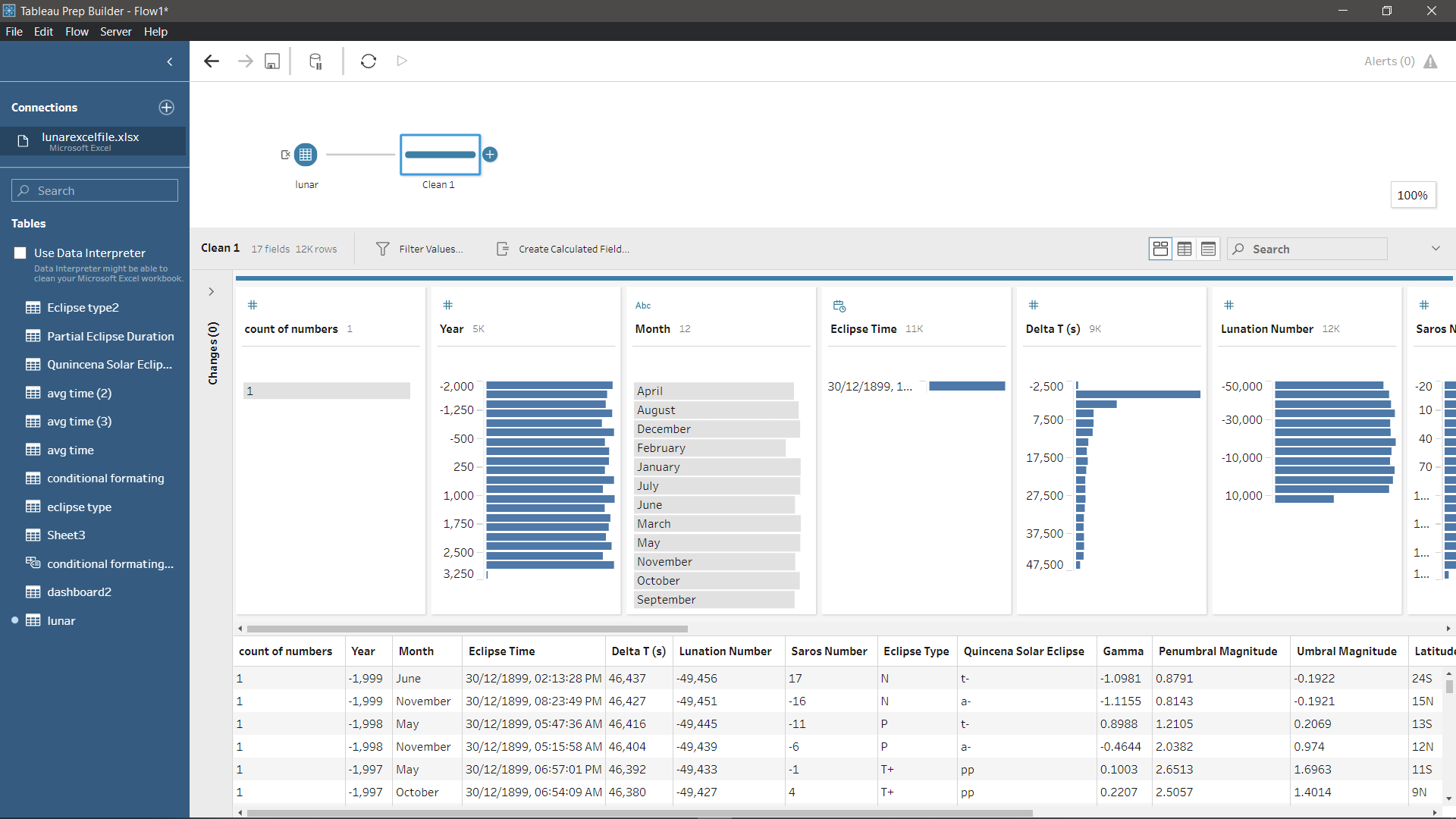
Loading the data and unselecting the useless columns which are not part of the calculations



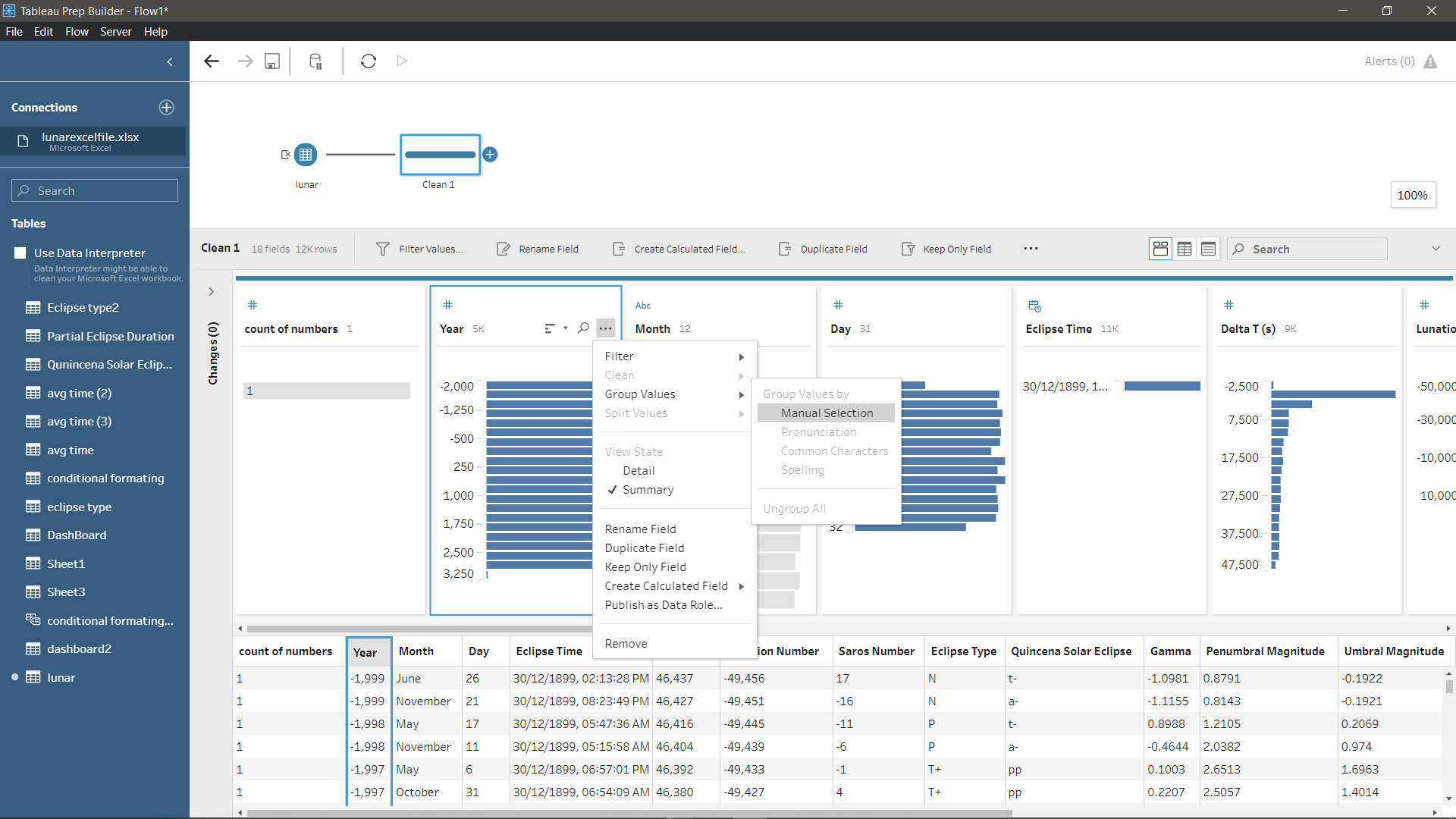
Selecting the clean step which is very important and crucial.



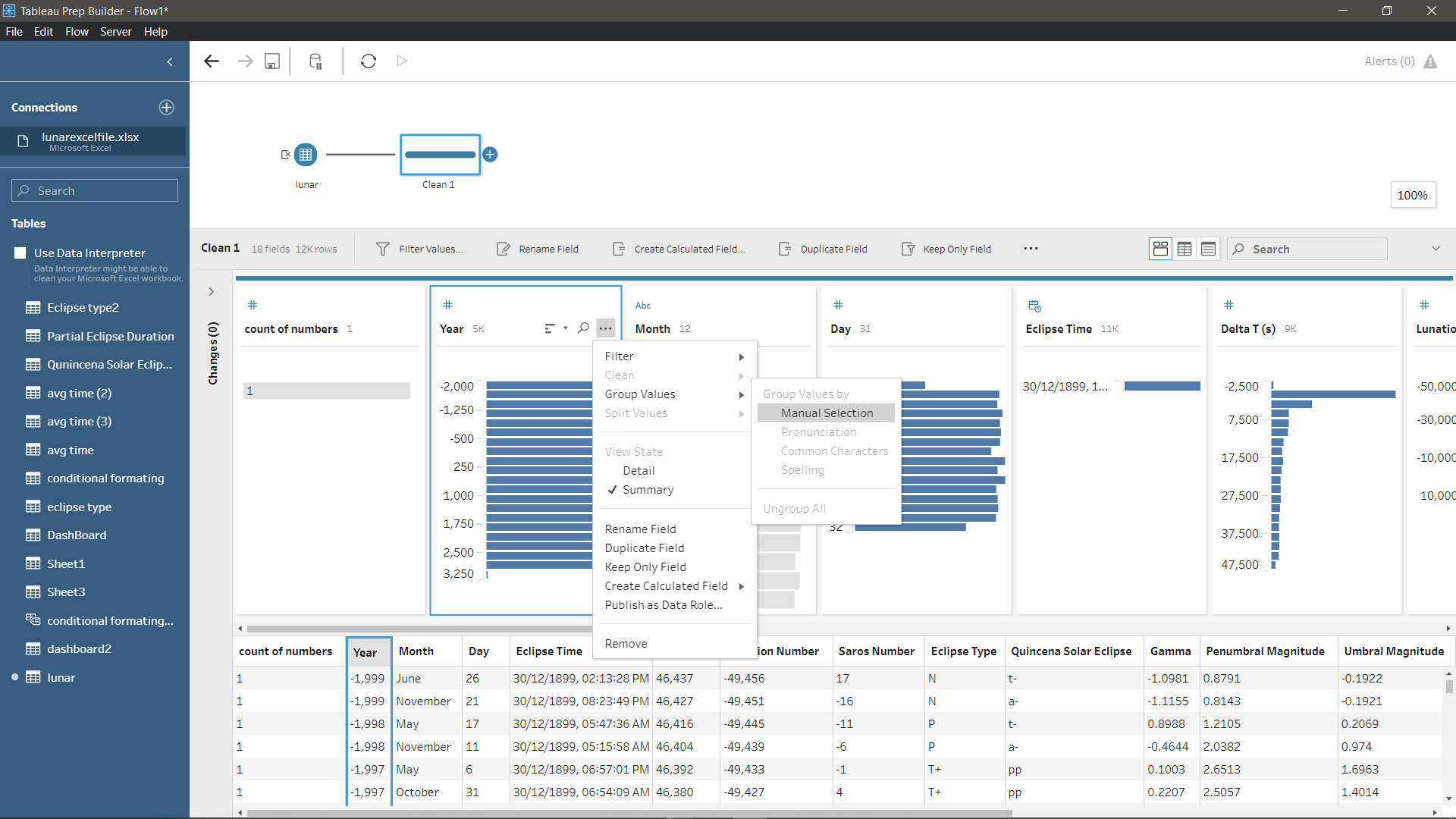
Now clean process is started and the summarized table is going to display as shown.



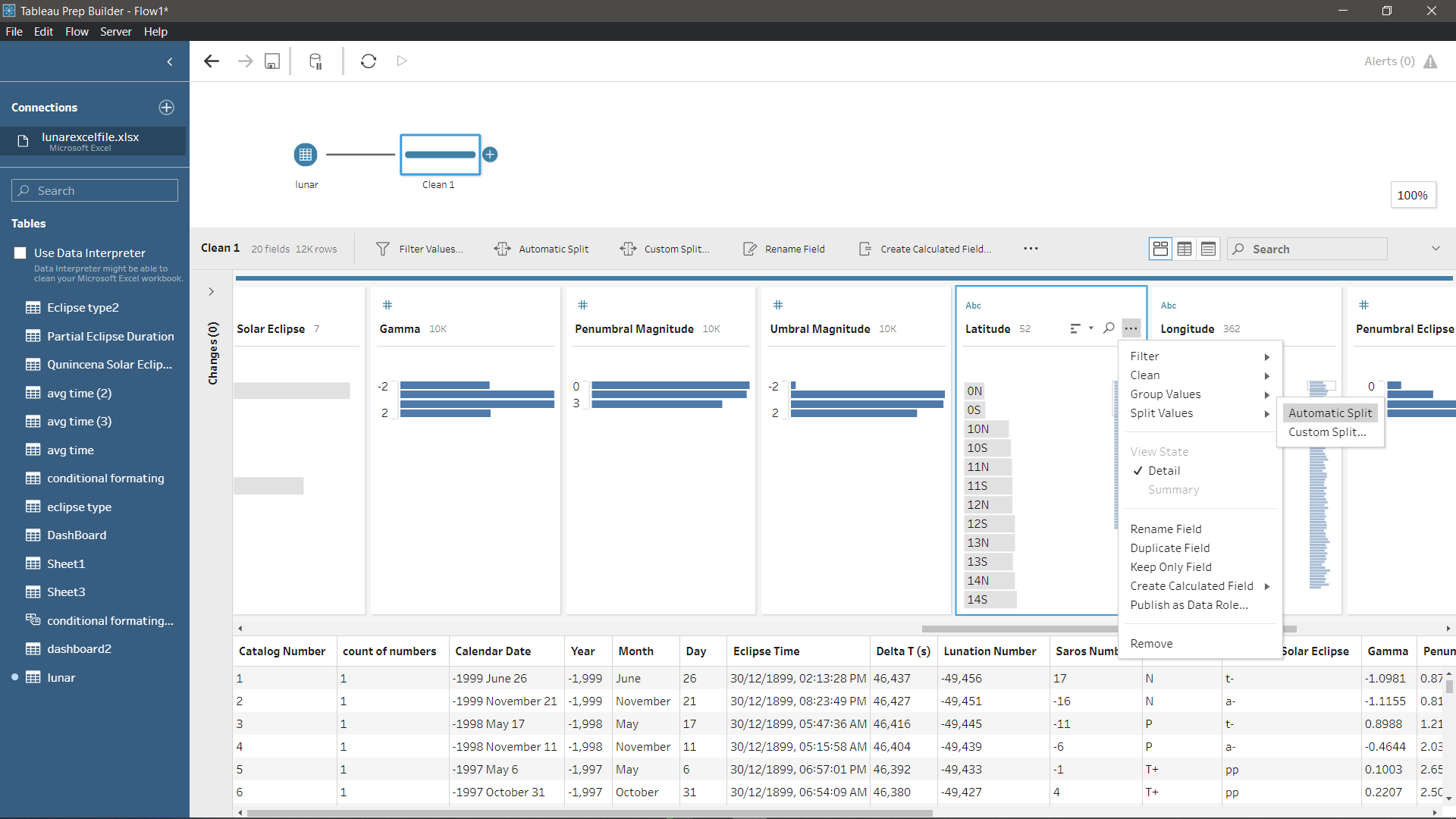
Now the process of grouping the cells ‘year’ so that easy view of the data and simplified view going to take place.



In Partial eclipse duration there are some null values which can be removed using the filter option and null values option.



In the ‘Eclipse Time’ the value of year and is given in this year is not useful so we gonna split and divide the time separately



**References:-**

<https://www.kaggle.com/nasa/solar-eclipses>

<http://www.cmbi.com.au/S011_PowerPivot_Sample_Workbook.html>

<https://www.tutorialspoint.com/excel_data_analysis/index.html>

<https://www.analyticsvidhya.com/blog/2020/04/excel-tips-tricks-data-analysis/>