University of Camerino

School of Science and Technology

Master Degree in Computer Science (LM-18)

Course of Data Analytics

Data analysis on YoUnicam App dataset - YOUNICAM#STATS

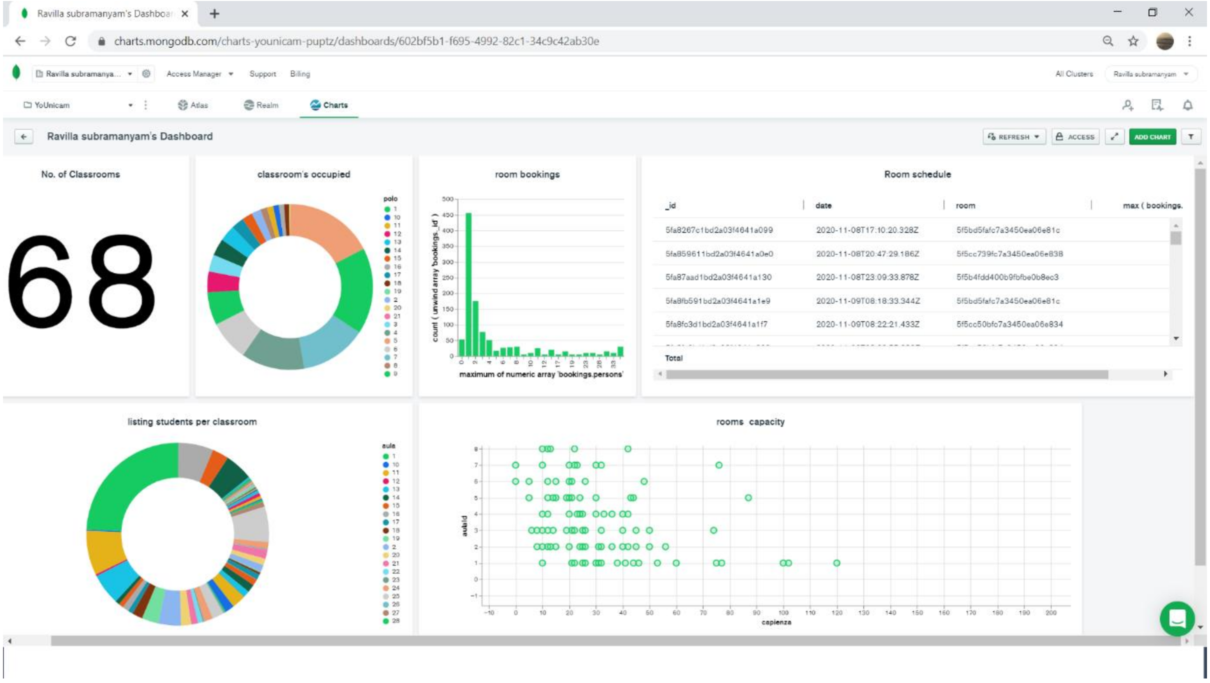
Students:
R.S. Keerthi
Ahmed Dodia

Professor
Dr.**Massimo Callisto De Donato**

A.A. 2020/2021

CONTENTS:

1 INTRODUCTION


Ravilla subramanyam's Dashboard
charts.mongodb.com/charts-younicam-puptz/dashboards/602bf5b1-f695-4992-82c1-34c9c42ab30e
Access Manager
Support
Billing
All Clusters
Ravilla subramanyam
YoUnicam
Atlas
Realm
Charts
REFRESH
ACCESS
ADD CHART
No. of Classrooms
68
classroom's occupied
polo
room bookings
count (unwind array bookings._id)
maximum of numeric array 'bookings.persons'
listing students per classroom
aula
rooms capacity
aubld
capienza

Room schedule

_id	date	room	max (bookings.
5fa8267c1bd2a03f4641a099	2020-11-08T17:10:20.328Z	5f5bd5fafc7a3450ea06e81c	
5fa859611bd2a03f4641a0e0	2020-11-08T20:47:29.186Z	5f5cc739fc7a3450ea06e838	
5fa87aad1bd2a03f4641a130	2020-11-08T23:09:33.878Z	5f5b4fdd400b9fbfbe0b8ec3	
5fa8fb591bd2a03f4641a1e9	2020-11-09T08:18:33.344Z	5f5bd5fafc7a3450ea06e81c	
5fa8fc3d1bd2a03f4641a1f7	2020-11-09T08:22:21.433Z	5f5cc50bfc7a3450ea06e834	
Total			

Data visualization has become more important for getting more insights on data. As one of the essential steps in the business intelligence process, data visualization takes the raw data, models it, and delivers the data so that conclusions can be reached. In advanced analytics, data scientists are creating machine learning algorithms to better compile essential data into visualizations that are easier to understand and interpret.

Specifically, data visualization uses visual data to communicate information in a manner that is universal, fast, and effective. This practice can help companies identify which areas need to be improved, which factors affect customer satisfaction and dissatisfaction, and what to do with specific products (where should they go and who should they be sold to). Visualized data gives stakeholders, business owners, and decision-makers a better prediction of sales volumes and future growth. We will be using the mongodb charts for visualizing the mongodb data. It provides features like Aggregation Functionality, Seamless Integration with MongoDB Atlas, Document Data Handling.

MongoDB Charts makes communicating your data a straightforward process by providing built-in tools to easily share and collaborate on visualizations. Here in these projects, we will be cleaning and analyzing the YoUnicam app data on the Mongodb charts. And we will be focusing on creating different viewpoints for the same data to provide more insight to the data.

2. REQUIREMENTS

In this section we are going to describe all the requirements of Mongodb charts:

MONGODB ATLAS ACCOUNT

Mongodb charts can be accessible through the mongodb atlas. You will need a mongodb atlas account to complete a procedure. If you're an MongoDb cloud manager user, you can use your Cloud Manager credentials but, to access an organization or a project with in organization, you must be a member of that organization.

To access clusters in a project, users must belong to that project. Users in MongoDB Atlas can belong to multiple projects within an organization, you can group that users into team. Users can belong to multiple teams. These teams further can be assigned to multiple projects, and team members to access the project determined by the team's project role.

Reference link for creating an account https://account.mongodb.com/account/login .

PROJECT ROLES

To access the MongoDB charts, Atlas users should have appropriate access to read the cluster data from the project. Atlas user can have any of these project roles like project owner, project cluster manager, project data access admin, project data access read/write, project data access read only except the project read only access as it doesn't grant the appropriate permissions. When we first register as an atlas user, we will be granted project owner role for the initial project by default.

ATLAS PROJECT

To create the project for an organization, you must be either an organization owner or organization project creator. You can create multiple projects within an organization. The benefits of these projects in organization are that you can isolate different environments from each other, you can maintain separate clusters with different security configurations, you can also assign different users and teams to different environments and give different permissions, and mainly you can set different alert settings.

Create an atlas project or if you already have an atlas project with clusters associated with it by having data that you want visualize, you can use that data as data source and create mongodb charts.

ATLAS CLUSTER

Clusters are Atlas-managed Mongodb deployments, A cluster can either be a replica set or a sharded cluster. The MongoDB charts application makes it easy to add collections for your cluster as data sources. Data sources reference specific collections which you can access in the Chart Builder to visualize the data in those collections.

DATA SOURCES

Data sources in MongoDB Charts reference a collection or view in your MongoDB deployment. The fields in that collection or view can be used to construct a chart. When building a chart, you will need to specify the data source that the chart uses. You have to provide the data that you need to visualize as data source which you would have imported using mongo shell or mongodb compass or dashboard.

3 TECHNOLOGIES USED

PANDAS

Pandas is an open-source python package that is most widely used for data science/data analysis and machine learning tasks. It is built on top of another package named Numpy, which provides support for multi-dimensional arrays. As one of the most popular data wrangling packages, Pandas works well with many other data science modules inside the Python ecosystem, and is typically included in every Python distribution. Pandas makes it simple to do many of the time consuming, repetitive tasks associated with working with data, including: Data cleansing, data fill, data normalization, merges and joins, data Visualization, statistical analysis, data inspection, loading and saving data and much more.
we have used pandas for loading and conversion of json files to csv file using the methods read_json() , to_csv(). To describe and checking the relationships between the filed data we used the panda methods describe() and info().

NUMPY

NumPy stands for Numerical Python and is one of the most useful scientific libraries in Python programming. It provides support for large multidimensional array objects and various tools to work with them. Various other libraries like Pandas, Matplotlib, and Scikit-learn are built on top of this amazing library. In Python we have lists that serve the purpose of arrays, but they are slow to process. NumPy aims to provide an array object that is up to 50x faster than traditional Python lists. The array object in NumPy is called ndarray, it provides a lot of supporting functions that make working with ndarray very easy. Arrays are very frequently used in data science, where speed and resources are very important.
NumPy arrays are stored at one continuous place in memory unlike lists, so processes can access and manipulate them very efficiently. This behaviour is called locality of reference in computer science. This is the main reason why NumPy is faster than lists. Also, it is optimized to work with latest CPU architectures.

We have used numpy for checking the uniqueness of the values in the data using method np.unique() and checking the null or empty values in the data using the method np,isnull() converting the types of data using the np.to_numeric().

JUPYTER NOTEBOOK

The Jupyter Notebook is an open-source web application that you can use to create and share documents that contain live code, equations, visualizations, and text. Jupyter Notebook is maintained by the people at Project Jupyter. These Jupyter notebook app is a server-client application that allows editing and running notebook documents via a web browser. The Jupyter Notebook App can be executed on a local desktop requiring no internet access (as described in this document) or can be installed on a remote server and accessed through the internet.
We have used jupyter notebook to write code for cleaning and formatting the giving data and

we have used jupyter for sharing the entire code that we have done to the github repository.

MONGO SHELL

MongoDB Mongo shell is an interactive JavaScript interface, that allows you to interact with MongoDB instances through the command line. It is the quickest way to connect, configure, query, and work with your MongoDB database. MongoDB Shell provides a modern command-line experience that includes syntax highlighting, intelligent autocomplete, contextual help, and clear error messages. These are just some of the features included in MongoDB Shell that makes working with your MongoDB Databases easier.

The MongoDB Shell is a standalone product, it's developed separately from the MongoDB Server and it's open-source under the Apache 2 license. We have used mongo shell to connect to our mongodb database and for importing and exporting the data from the database using the commands.

MONGODB COMPASS

As the GUI for MongoDB, MongoDB Compass allows you to make smarter decisions about document structure, querying, indexing, document validation, and more. Commercial subscriptions include technical support for MongoDB Compass. MongoDB Compass analyses your documents and displays rich structures within your collections through an intuitive GUI. It allows you to quickly visualize and explore your schema to understand the frequency, types and ranges of fields in your data set. The Compass Plugin Framework is exposed as an API, making it extensible by users. Looking for other functionality? Install a plugin or build your own.

MongoDB Compass makes aggregation easier. By constructing aggregation pipelines in an intuitive UI. Code skeletons and auto-complete make it easy to build a stage, while a preview of documents shows you if the stage is doing what you need. Add stages, remove them, or drag and drop to re-order in the pipeline. Once you're done, export it to native code to use in your application.

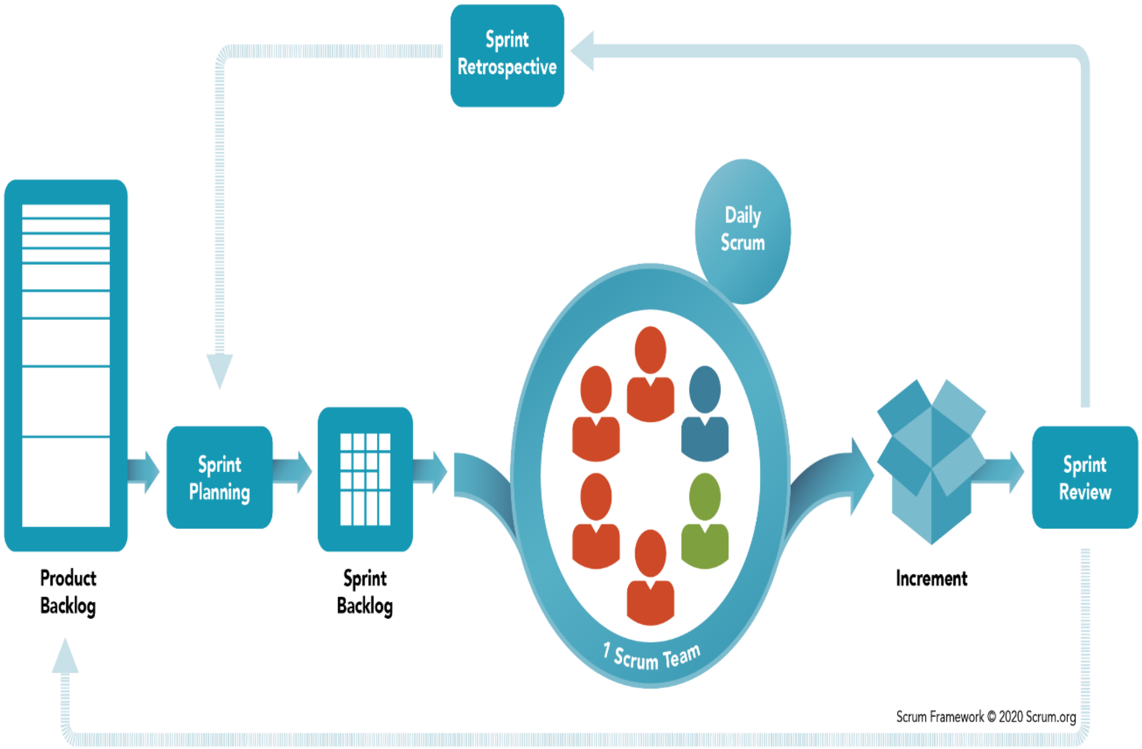
4 METHODOLOGY

SCRUM

Scrum is an agile way to manage a project, usually software development. Agile software development with Scrum is often perceived as a methodology; but rather than viewing Scrum as methodology, think of it as a framework for managing a process.

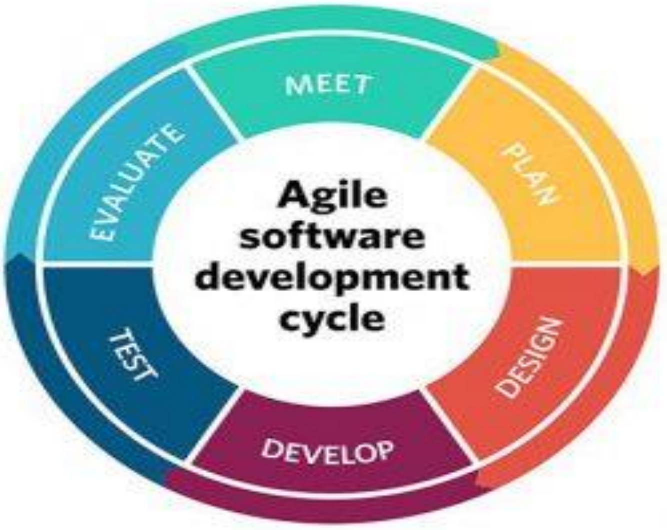
Scrum is a lightweight framework that helps people, teams and organizations generate value through adaptive solutions for complex problems. The three roles defined in Scrum are the ScrumMaster, the Product Owner, and the Team (which consists of Team members). The people who fulfill these roles work together closely, on a daily basis, to ensure the smooth flow of information and the quick resolution of issues.

Scrum is simple. It is the opposite of a big collection of interwoven mandatory components. Scrum is not a methodology. Scrum implements the scientific *method* of empiricism. Scrum replaces a programmed algorithmic approach with a heuristic one, with respect for people and self-organization to deal with unpredictability and solving complex problems. The below graphic represents Scrum in Action as described by Ken Schwaber and Jeff Sutherland in their book *Software in 30 Days* taking us from planning through software delivery.



AGILE

Agile software development refers to software development methodologies centered round the idea of iterative development, where requirements and solutions evolve through collaboration between self-organizing cross-functional teams. The ultimate value in Agile development is that it enables teams to deliver value faster, with greater quality and predictability, and greater aptitude to respond to change. Scrum and Kanban are two of the most widely used Agile methodologies. Below are the most frequently asked questions around Agile and Scrum, answered by our experts.



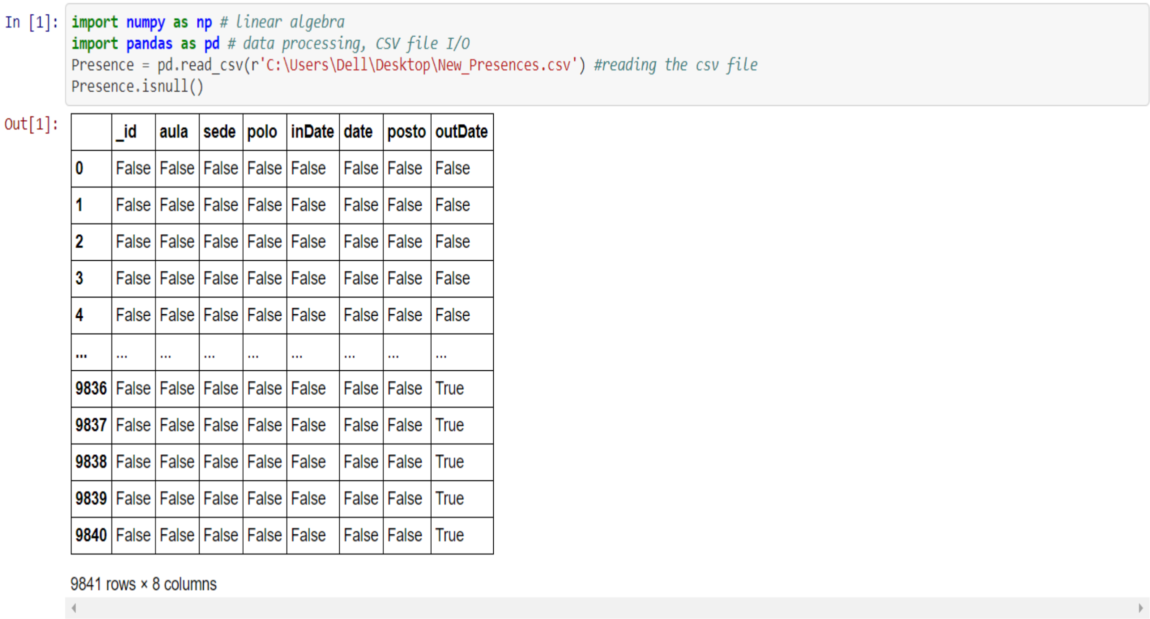
5 OBJECTIVES

The main objective is to clean the collected data, analysis it and providing statistics on the MongoDB charts.

CLEANING THE DATA

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled. We have used the juypter notebook to work on the data and we also used some methods to make data free from the irrelevant data.

1. **Handling the missing data:**
   We have checked the data frames for the missing fields and null values. If any existed, we have looked for the importance of that filed to the other data, if the filed is important we have replaced the missing data with the average of that fields data or some times with some random common data.



In [1]:

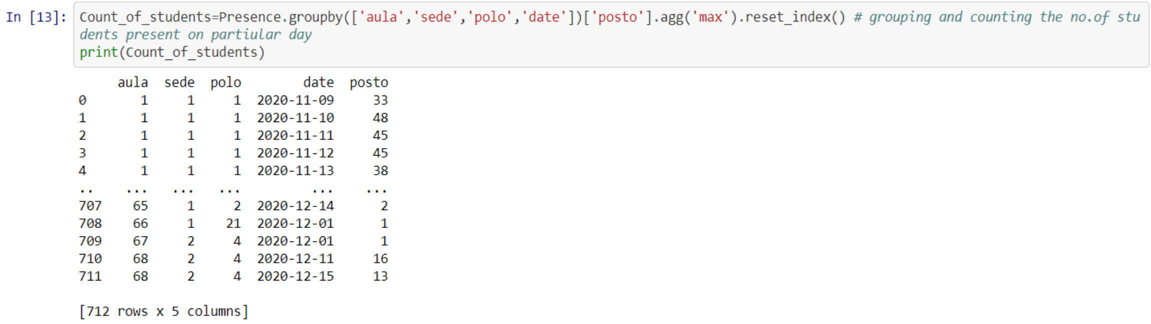
```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O
Presence = pd.read_csv(r'C:\Users\Dell\Desktop\New_Presences.csv') #reading the csv file
Presence.isnull()
```

Out[1]:

	_id	aula	sede	polo	inDate	date	posto	outDate
0	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...
9836	False	False	False	False	False	False	False	True
9837	False	False	False	False	False	False	False	True
9838	False	False	False	False	False	False	False	True
9839	False	False	False	False	False	False	False	True
9840	False	False	False	False	False	False	False	True

9841 rows × 8 columns

2. **Irrelevant data :**
   Irrelevant observations are those that don't actually fit the specific problem that your trying to solve. That can be said as extra data that has no impact on the data frames or can be used for analysis.

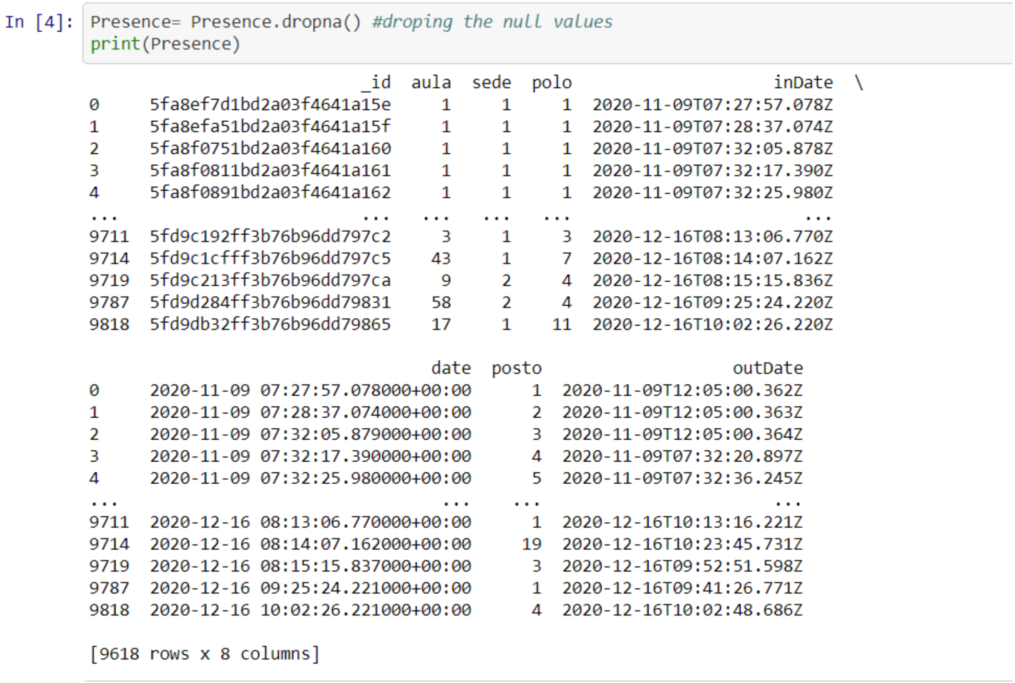


```
In [13]: Count_of_students=Presence.groupby(['aula','sede','polo','date'])['posto'].agg('max').reset_index() # grouping and counting the no.of stu
dents present on partiular day
print(Count_of_students)

     aula  sede  polo        date  posto
0       1     1     1  2020-11-09     33
1       1     1     1  2020-11-10     48
2       1     1     1  2020-11-11     45
3       1     1     1  2020-11-12     45
4       1     1     1  2020-11-13     38
..    ...   ...   ...         ...    ...
707    65     1     2  2020-12-14      2
708    66     1    21  2020-12-01      1
709    67     2     4  2020-12-01      1
710    68     2     4  2020-12-11     16
711    68     2     4  2020-12-15     13

[712 rows x 5 columns]
```

3. **Dropping the null values**:
   If the filed that has null values has no major effect on the remaining data. It is better to drop it to reduce the complexity.



```
In [4]: Presence= Presence.dropna() #droping the null values
print(Presence)

                           _id  aula  sede  polo                     inDate  \
0     5fa8ef7d1bd2a03f4641a15e     1     1     1  2020-11-09T07:27:57.078Z
1     5fa8efa51bd2a03f4641a15f     1     1     1  2020-11-09T07:28:37.074Z
2     5fa8f0751bd2a03f4641a160     1     1     1  2020-11-09T07:32:05.878Z
3     5fa8f0811bd2a03f4641a161     1     1     1  2020-11-09T07:32:17.390Z
4     5fa8f0891bd2a03f4641a162     1     1     1  2020-11-09T07:32:25.980Z
...                        ...   ...   ...   ...                       ...
9711  5fd9c192ff3b76b96dd797c2     3     1     3  2020-12-16T08:13:06.770Z
9714  5fd9c1cfff3b76b96dd797c5    43     1     7  2020-12-16T08:14:07.162Z
9719  5fd9c213ff3b76b96dd797ca     9     2     4  2020-12-16T08:15:15.836Z
9787  5fd9d284ff3b76b96dd79831    58     2     4  2020-12-16T09:25:24.220Z
9818  5fd9db32ff3b76b96dd79865    17     1    11  2020-12-16T10:02:26.220Z

                                 date  posto                   outDate
0     2020-11-09 07:27:57.078000+00:00      1  2020-11-09T12:05:00.362Z
1     2020-11-09 07:28:37.074000+00:00      2  2020-11-09T12:05:00.363Z
2     2020-11-09 07:32:05.879000+00:00      3  2020-11-09T12:05:00.364Z
3     2020-11-09 07:32:17.390000+00:00      4  2020-11-09T07:32:20.897Z
4     2020-11-09 07:32:25.980000+00:00      5  2020-11-09T07:32:36.245Z
...                               ...    ...                       ...
9711  2020-12-16 08:13:06.770000+00:00      1  2020-12-16T10:13:16.221Z
9714  2020-12-16 08:14:07.162000+00:00     19  2020-12-16T10:23:45.731Z
9719  2020-12-16 08:15:15.837000+00:00      3  2020-12-16T09:52:51.598Z
9787  2020-12-16 09:25:24.221000+00:00      1  2020-12-16T09:41:26.771Z
9818  2020-12-16 10:02:26.221000+00:00      4  2020-12-16T10:02:48.686Z

[9618 rows x 8 columns]
```

4. **Filter unwanted outliers**:
   Outliers can cause problems with certain types of models. For example, linear regression models are less robust to outliers than decision tree models. In general, if you have a legitimate reason to remove an outlier, it will help your model's performance. However, outliers are innocent until proven guilty. You should never remove an outlier just because it's a "big number." That big number could be very informative for your model. We can't stress this enough: you must have a good reason for removing an outlier, such as suspicious measurements that are unlikely to be real data.

Data Analysis:

Data can be analysed using some standard and important statistical methods they are as follows:

1. MEAN:

The arithmetic mean, more commonly known as "the average," is the sum of a list of numbers divided by the number of items on the list. The mean is useful in determining the overall trend of a data set or providing a rapid snapshot of your data. Another advantage of the mean is that it's very easy and quick to calculate.

Pitfall:

Taken alone, the mean is a dangerous tool. In some data sets, the mean is also closely related to the mode and the median (two other measurements near the average). However, in a data set with a high number of outliers or a skewed distribution, the mean simply doesn't provide the accuracy you need for a nuanced decision.

2. STANDARD DEVIATION

The standard deviation, often represented with the Greek letter sigma, is the measure of a spread of data around the mean. A high standard deviation signifies that data is spread more widely from the mean, where a low standard deviation signals that more data align with the mean. In a portfolio of data analysis methods, the standard deviation is useful for quickly determining dispersion of data points.

Pitfall:

Just like the mean, the standard deviation is deceptive if taken alone. For example, if the data have a very strange pattern such as a non-normal curve or a large amount of outliers, then the standard deviation won't give you all the information you need.

3. REGRESSION

Regression models the relationships between dependent and explanatory variables, which are usually charted on a scatterplot. The regression line also designates whether those relationships are strong or weak. Regression is commonly taught in high school or college statistics courses with applications for science or business in determining trends over time.

Pitfall:

Regression is not very nuanced. Sometimes, the outliers on a scatterplot (and the reasons for them) matter significantly. For example, an outlying data point may represent the input from your most critical supplier or your highest selling product. The nature of a regression line, however, tempts you to ignore these outliers.

4. SAMPLE SIZE DETERMINATION

When measuring a large data set or population, like a workforce, you don't always need to collect information from every member of that population – a sample does the job just as well. The trick is to determine the right size for a sample to be accurate. Using proportion and standard deviation methods, you are able to accurately determine the right sample size you need to make your data collection statistically significant.

Pitfall:

When studying a new, untested variable in a population, your proportion equations might need to rely on certain assumptions. However, these assumptions might be completely inaccurate. This error is then passed along to your sample size determination and then onto the rest of your statistical data analysis

5. HYPOTHESIS TESTING

Also commonly called *t* testing, hypothesis testing assesses if a certain premise is actually true for your data set or population. In data analysis and statistics, you consider the result of a hypothesis test *statistically significant* if the results couldn't have happened by random chance. Hypothesis tests are used in everything from science and research to business and economic

Pitfall:

To be rigorous, hypothesis tests need to watch out for common errors. For example, the placebo effect occurs when participants falsely expect a certain result and then perceive (or actually attain) that result. Another common error is the Hawthorne effect (or observer effect), which happens when participants skew results because they know they are being studied. Overall, these methods of data analysis add a lot of insight to your decision making portfolio, particularly if you've never analysed a process or data set with statistics before. However, avoiding the common pitfalls associated with each method is just as important. Once you master these fundamental techniques for statistical data analysis, then you're ready to advance to more powerful data analysis tools.

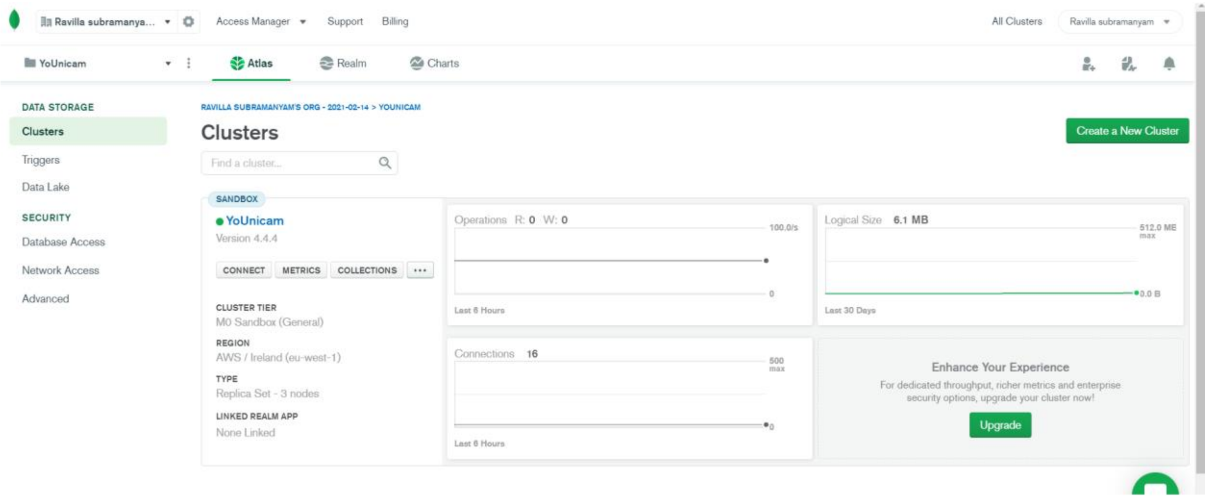
MongoDB Charts:

MongoDB Charts is accessible through MongoDB Atlas. You will need a MongoDB Atlas user account to complete this procedure. The procedure to create a mongodb chart is as follows.

1. Log into MongoDB Atlas: To access the MongoDB Charts application, you must be logged into Atlas.

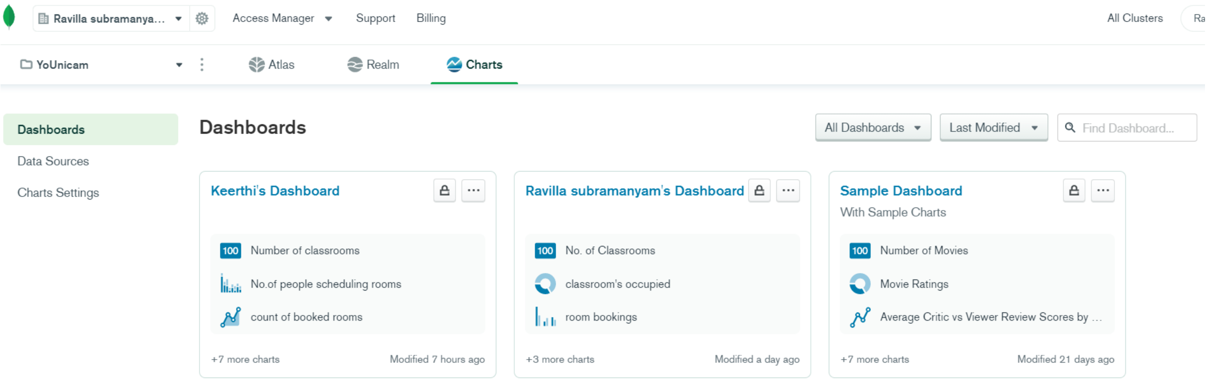
2. Select your desired Atlas project, or create a new project: If you have an Atlas project with clusters containing data you wish to visualize, select the project from the Context dropdown in the left navigation pane.

3. Create an Atlas cluster: The MongoDB Charts application makes it easy to add collections in your cluster as data sources. Data sources reference specific collections which you can access in the Chart Builder to visualize the data in those collections.



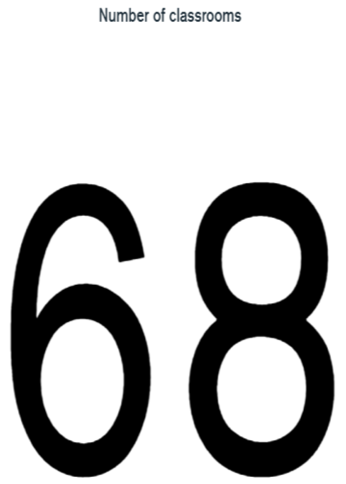
4. Launch the MongoDB Charts application:

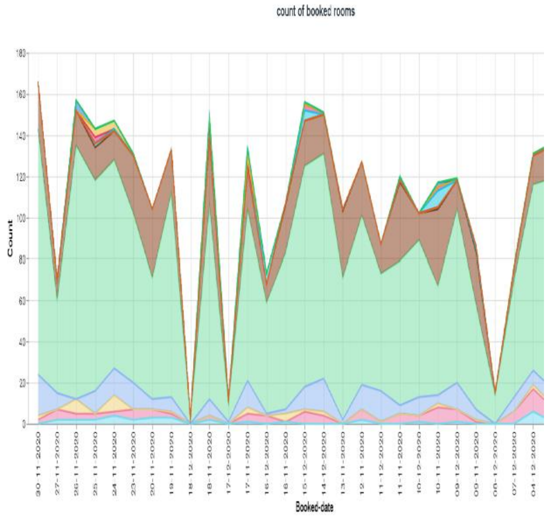
If you are a new Charts user, MongoDB Charts directs you to the Charts Welcome Experience. The Charts Welcome Experience provides two possible paths to begin using Charts either by connecting your data source or by explore with sample data. Additionally, MongoDB Charts automatically creates a new, empty dashboard named <YOUR-NAME>'s Dashboard of which you are the Owner. This dashboard is private by default, but you can modify dashboard permissions as you would any other dashboard.

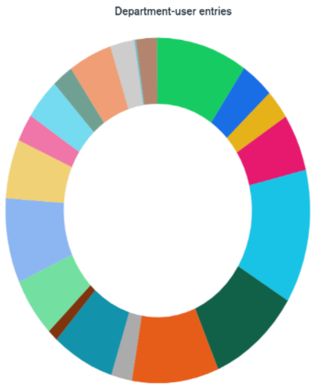


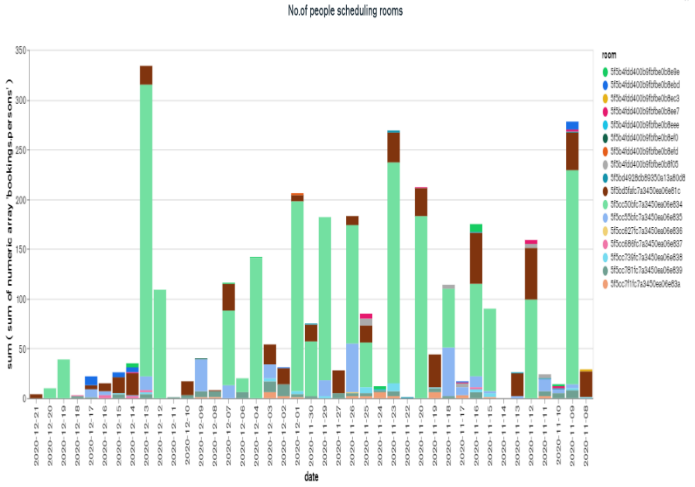
5 ACHIEVED RESULTS:

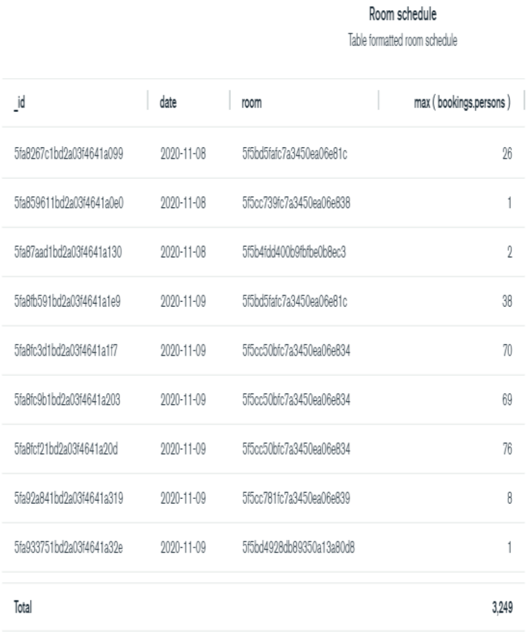
The easily understandable visualized data in the form of charts using the MongoDB charts as the following for the YoUnicam app data.







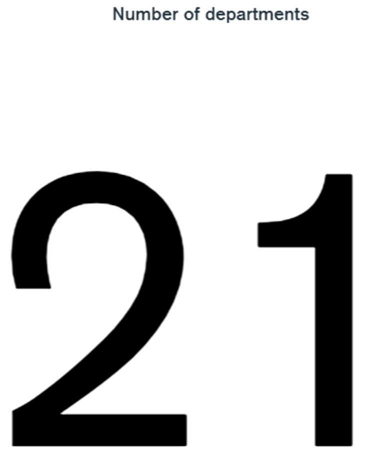




Room schedule

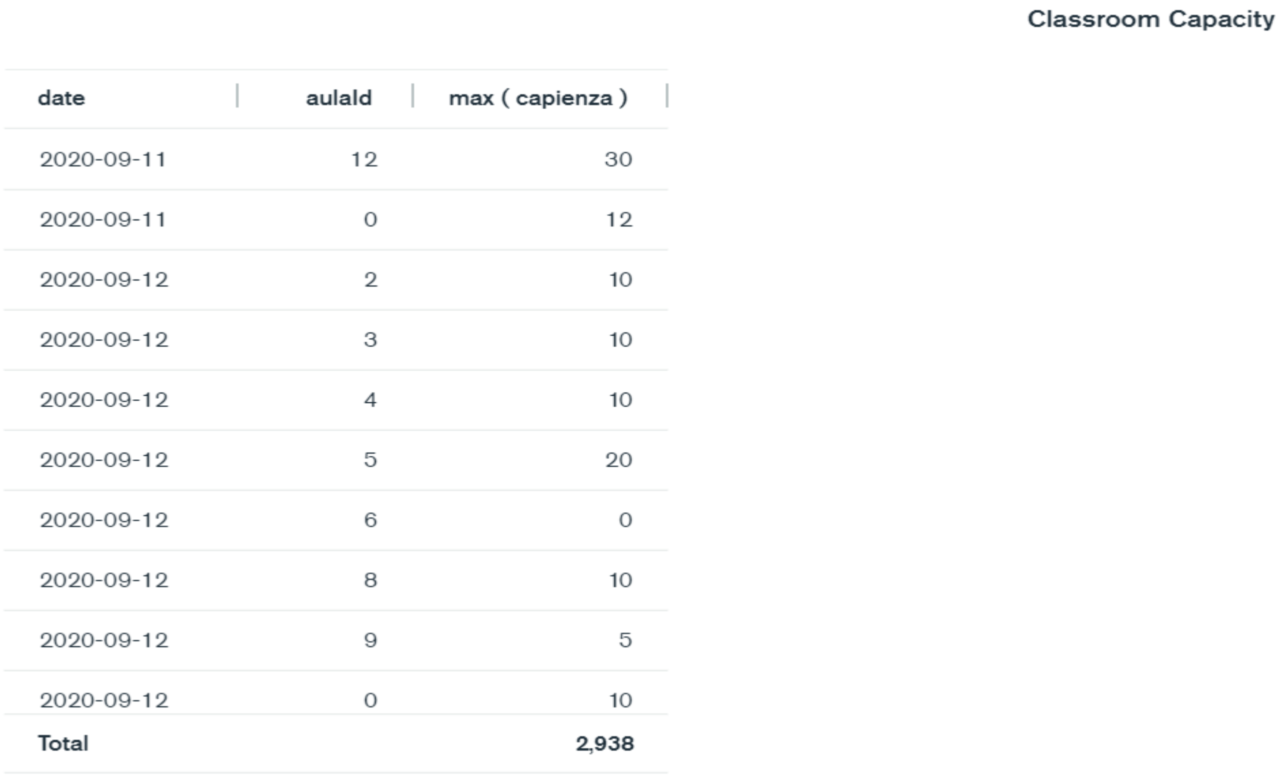
Table formatted room schedule

_id	date	room	max (bookings.persons)
5fa8267c1bd2a03f4641a099	2020-11-08	5f5bd5fafc7a3450ea06e81c	26
5fa859611bd2a03f4641a0e0	2020-11-08	5f5cc739fc7a3450ea06e838	1
5fa87aad1bd2a03f4641a130	2020-11-08	5f5b4fdd400b9fbfbe0b8ec3	2
5fa8fb591bd2a03f4641a1e9	2020-11-09	5f5bd5fafc7a3450ea06e81c	38
5fa8fc3d1bd2a03f4641a1f7	2020-11-09	5f5cc50bfc7a3450ea06e834	70
5fa8fc9b1bd2a03f4641a203	2020-11-09	5f5cc50bfc7a3450ea06e834	69
5fa8fcf21bd2a03f4641a20d	2020-11-09	5f5cc50bfc7a3450ea06e834	76
5fa92a841bd2a03f4641a319	2020-11-09	5f5cc781fc7a3450ea06e839	8
5fa933751bd2a03f4641a32e	2020-11-09	5f5bd4928db89350a13a80d8	1
Total			3,249



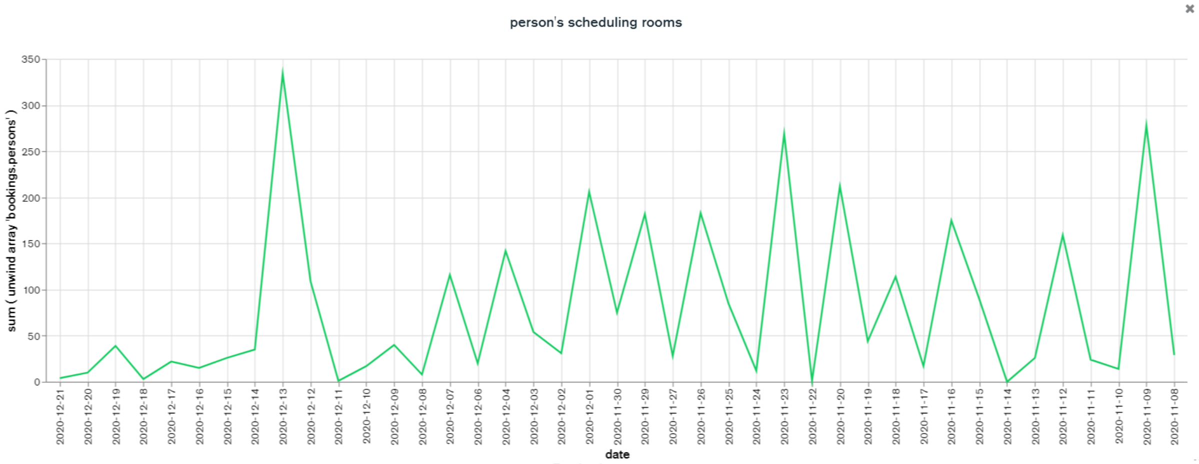
Number of departments

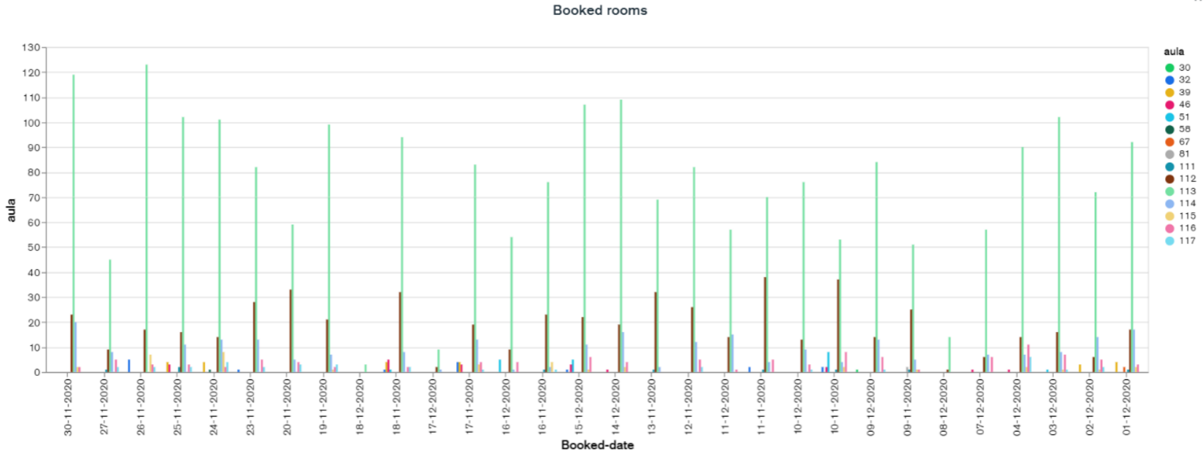
21



Classroom Capacity

date	auladId	max (capienza)
2020-09-11	12	30
2020-09-11	0	12
2020-09-12	2	10
2020-09-12	3	10
2020-09-12	4	10
2020-09-12	5	20
2020-09-12	6	0
2020-09-12	8	10
2020-09-12	9	5
2020-09-12	0	10
Total		2,938


person's scheduling rooms
sum (unwind array 'bookings.persons')
350
300
250
200
150
100
50
0
2020-12-21
2020-12-20
2020-12-19
2020-12-18
2020-12-17
2020-12-16
2020-12-15
2020-12-14
2020-12-13
2020-12-12
2020-12-11
2020-12-10
2020-12-09
2020-12-08
2020-12-07
2020-12-06
2020-12-04
2020-12-03
2020-12-02
2020-12-01
2020-11-30
2020-11-29
2020-11-27
2020-11-26
2020-11-25
2020-11-24
2020-11-23
2020-11-22
2020-11-20
2020-11-19
2020-11-18
2020-11-17
2020-11-16
2020-11-15
2020-11-14
2020-11-13
2020-11-12
2020-11-11
2020-11-10
2020-11-09
2020-11-08
date


Booked rooms
aula
130
120
110
100
90
80
70
60
50
40
30
20
10
0
30-11-2020
27-11-2020
26-11-2020
25-11-2020
24-11-2020
23-11-2020
20-11-2020
19-11-2020
18-12-2020
18-11-2020
17-12-2020
17-11-2020
16-12-2020
16-11-2020
15-12-2020
14-12-2020
13-11-2020
12-11-2020
11-12-2020
11-11-2020
10-12-2020
10-11-2020
09-12-2020
09-11-2020
08-12-2020
07-12-2020
04-12-2020
03-12-2020
02-12-2020
01-12-2020
Booked-date
aula
30
32
39
46
51
58
67
81
111
112
113
114
115
116
117