**NAME**

**COLLEGE NUMBER**

**DATA DESIGN AND DEVELOPMENT: A BI APPROACH**

**Abstract**

The development of database systems should be considered and looked at as a wholesome approach. Since data is the heart of the organisation, organizations and individuals are working their way up the ladder to come up with more refined ways of offering information to management. There are three processes to consider when trying to deal with the data system. They include collection, storage, manipulation and visualisation.

**Data collection:**

Is considered the first process in decision making. Top management relies on decision support tools to come up with ready to use algorithms and tools that will necessitate the process of highlighting key decisions to be taken in different methods of addressing the company issues. Sources of these data can also be referred to as data points. Data points are key locations where data gets into the storage devices.

Usually, the company has servers within the applications upon its network that are meant for the consumption and full use of the information. Examples of such data points include. Customer relationship arrangements, company mobile apps, website portals, points of sales, contact forms and link, call logs, email conversations and social media responses. Now what happens in the event that the company is receiving such too much information that it may not be able to handle, this is where big data now comes in. Big data will try to answer the following questions;

1. What type of data is this?
2. Where is it coming from?
3. Who is sending it in?
4. How should we store it?
5. How can we analyse it?

Some of the features of big data include the fact that it’s fast moving, it’s in high amounts and it contains a larger amounts of truth in it.

In the next step of the data process includes the data storage. Also, some kinds of online data storage equally occur and leaves traces of data along the path. Instance situation is where you go shopping and then you swipe your card across the POS, card reader or o some machine, the traces of you r information is left all over. Alternatively, when we also go online to search for a particular product to buy, Google search engine automatically and secretly monitors our key words, maps these to our locations and then use these information to send to use recommendations from other sites based on the type of item and the want on what to buy and how closely its related to the items we are actually looking for online.

This is one of the reasons why sometimes after searching for a particular product on, you get so many recommendations from other sites, mobile apps and social media pages that you visit. This kind of targeted selling may appear illegal and uncalled for but in some way tots still allowed. May be we need to review further the kind of laws and policies that we put in place to safeguard the human data privacy, such privileges should d not just be protected by some weird terms and conditions that usually consumers do not even read or understand.

Further looking at data storage, we can consider that applications majorly store data on three different types of formats.

* Structured
* Unstructured
* No storage (web 3.0)

**Structured database systems** is the kind of traditional database systems that is basically stored in the form of organised rows and columns, the attention here is paid to the organised rows that contains the unique primary and secondary keys that are used to access these databases. These data values are stored and exists table. A table could contain several columns for storing different tables names also known as variables. A variable is a storage name for the particular type of data. They can store strings, numbers also known as integers, characters, Booleans and even float variables.

The data is managed by the language known as the structured query language also known as SQL the SQL runs by requesting formation stored on several tables and columns. It can also give database information, schemas, metadata, and table information and then join different tables together as we shall see later in this discussion. A database management system (DBMS), on the other hand is as a tool for managing structured data. Examples of these DBMS include the MySQL, Oracle, DB2 and Postgres.

Almost all these databases work in the same way, the difference however could be a little syntax agreements here and there, the company name and the weather or not to accept capital SQL queries or just to ignore. Postgress accepts lower case written queries. Any other query that is written in upper case will automatically be converted back to lower case. The database is case sensitive.

Consider the below table sample for organised /structured database system

|  |  |  |  |
| --- | --- | --- | --- |
| Student\_id | name | wing | grade |
| 0003 | Alex maercies | Blue | 4 |
| 773 | Jontahan Mario | Red | 3 |
| 6443 | Simon Trucey | Green | 1 |
| 5243 | Lucia Degraada | Yellow | 2 |
| 1109 | Kimberly Golden | Purple | 3 |

Structured databases systems have been used for period’s ad and are the oldest in history considering their wide range of use. They are the legacy systems and databases. Their use and span from financial systems, healthcare education systems, research and development studies and in the financial stocks too, among other different developer environments.

Unstructured database systems on the other hand are the exact opposite of the structured database, Instead of data being stored in organised rows and columns, the data is tied in some unstructured format, meaning one cannot exactly tell if the data is in which particular row or index value position.

Unstructured data is the second latest type of database system, to get into the industry. Even some developers and software engineers are not well versed with the database. Sad but it’s true that this is where most applicators generating large amounts of data are going to shift their attention and focus to, this ideally true since UDBS is the only way the current applications tie with big data. Since UBD does not consider any special kind of specific rows and columns, the UDBS, will store images, audio, characters, integers and even documents that can only be accessed as an array. Consider the given example of an array calling the object below:

for (i in siteInfo .users)

{

    for (j in siteInfo.users[i])

    {

        x = siteInfo.users[i][j];

        console.log(x);

    }

}

The array above retrieved from unstructured database fetches an object called site info and returns the associated array of items within the object. This is the way this type of database works.

Google has one particular product called Firebase. This is tool that gives developer and organisational access to so many features including the Firebase data tools. This type of database does not define data in any kind of row or columns. Another type of database in use is the MongoDB, widely used for the same ,reasons as the Firebase .Also, the latest introduction to data storage winch in essence is not even stored is the web 3.0. You see when the internet was invented, it all began from static site to dynamic systems, and now to smart systems and now the whole system is now almost going back to web 1,0, but in a smart and silent way.

Instead of manually storing data in some scattered or unstructured database systems, the data now creates and destroys itself, users are identified by sessions and identifications. This is what crypto currency now trades in. No one knows the database or the particular data that is inside those applications. Instead, what we have are smart contracts. Smart contracts are barely owned various partners who run the etherium grid and every computer on the networks identified by key id of a particular user, when the user wants trade or buy ether, the particular smart contract are run on the block chain is called, stored in the system temporarily until l the transaction is marked as complete then contact is destroyed. Nowhere does the smart contact keep user details. This is for the same reasons why we say that block chain technology is safe and secure for transactions, it’s also safe in the long run. Can be it hacked? No.

The next step in the data process is the data analysis and storage. Developers and data engineers have come up with several tools, algorithms and languages that will sweeten the life of a data scientist/ scholar/ application engineer by making it easier to analyse data very fast. You see traditionally, after the introduction of the common business oriented language COBOL, introduced the COBOL tool in the 1960s, other scientists in the same domain have introduced scientific analysis tools like FORTAN and SCALA that focused on analysing data in the field of geography science.

Later on, developers would take into advantage the missing opportunity that was there for the engineers and thereafter to leverage on the needs of getting calculations and arithmetic applications on their dataset. This where the development of statistical languages such as R came in. R has been a powerful child in trying to help gat quick summary statistics on the concerned measures of central tendencies, summaries on the values of comparison and also by helping give the required forecasting and the data descriptive features and values of such datasets. Moreover the language has inbuilt mode so that will enable the researcher to quickly get whatever kind of insights that they need from such dataset. Some the models applied in this language include:

* Linear regressions
* Logistic regressions
* ARIMA models
* KNN model
* Random forest and decision trees

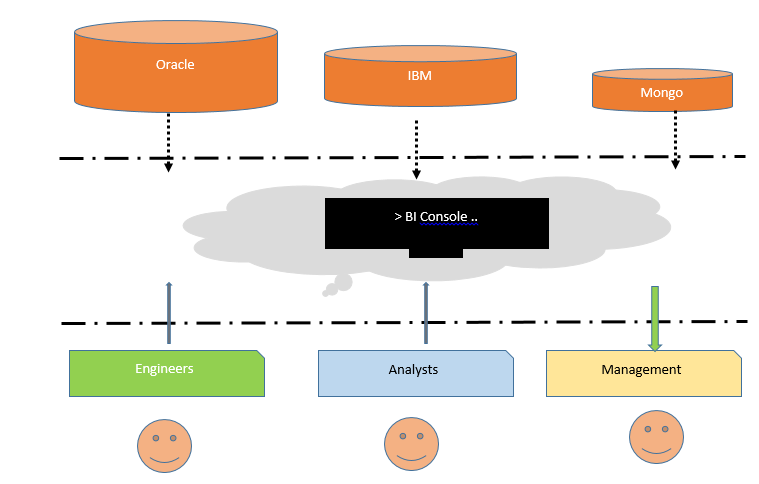
The models have been effective in trying to help scientists discover the relationships between datasets and still be able to predict the future outcomes of such inputs by forecasting the models through a fit function. Even though old and powerful, Python has been used to wrangle data during the whole data cleaning process, analysis and even building models to predict the dataset. Data analysis and features of this application include the ability to gather data from the internet through certain custom tools like beautiful soup. Once his data collection and preparation is ready then Pythons inbuilt libraries like Numpy, SciePy and pandas have been used as tools in trying to make the data visualisation process much easier.

Nonetheless, an old but still widely used approach to data analysis is a tool that we all know, Microsoft Excel is a tool that benefits almost every user and of basic analyst regardless of the experience or age of the researcher. Using Microsoft office, as we shall see in the latter chapters to come, an individual is able to quickly clean their dataset, get measures of central tendencies and then apply the relevant required functions on the data analyses tool pack to quickly get meaningful information from the dataset, as we shall discuss in the last part of this research papers in the analyses part..

Software as a service (SAS) are web tools that have been specifically designed to help the user quickly get around their analysis. With key attention to the analyses of the products and outputs derived from the dataset that they intend to use. Moreover, in this particular study the SAS tools shall form the basis of our analysis with key objectives and instructions being sent on how we can use the SAS software to convert a the CSV dataset of our choice into the correct .bat file that we can use on our application to make meaningful and useful data derivations from. Through a click and drag, we shall be quickly able to get the right analysis that is needed for this work.

All these introduction is necessary order to help the company to understand and measure the requirements of developing an effective business information system with a closer attention being paid to the datasets, systems and whether to consume the structured or unstructured database, the method and language of analysis and the kind of models to apply for each dataset. Finally, the two tools considered here for the development and management of the BI application shall be the SAS online tool and the earlier mentioned SQL language to help management understand the features and capabilities of BI applications.

**Introducing the BI Application**



The BI is kind of a three phased program. The upper phase that contains the databases needed for data storage, the middles phase where algorithms and formulas and functions plus models are stored and used and the lower layer that interacts with the BI process. The Bi in this case is considered to be stored in the cloud sever somewhere where engineers occasionally write and commit code to, maintenance and frequently scale up the application process.

The lower phase also contains the analysis who add custom queries and models on the dataset, rewrite program logics and execute per ad hoc business requirements to top management. Finally, the managements role in the whole BI thing is to get real time, reports attached click of a button. These reports can be downloaded in the form of PDF documents, XML charts or Microsoft Excel sheets. Management are the top targets in any BI system. The tool should act as a decision support material for them.

**Significance of BI systems**

The value of BI systems are not fetched. It has been established that organisations are able to increase their incomes and revenue streams by close to 30% according to a research done by KPMG in 2018. The reason why this would be so is attributed to the fact that BI systems give the organisation a 360 degree view of what is happening in the company, managers from various departments can monitor data and statistical performances on sales, customer acquisition, geographical mappings of stores and clients and how these regions are doing sales wise.

This kind and level of information is important to any manager as it will help them understand and make key decision on how to manage the products across the various regions of interest.

As a mechanism for sealing and patching revenue leaks, BI as a tool has aided in this. Several industry players lose millions of dollars since they do not have access on to who is using the products and how much they should actually be charged, but with the aid of a BI system, this is easier to realise and achieve.

Bi has also reduced the reporting time and key decisions can now be made faster. Top managers do not have to wait for analysts to extract data from servers, spend several minutes and hours doing cleaning and analysing, but instead, the whole analysis process is managed by the smart BI tool, projects inns are done and where necessary, KPI indicators are monitored ahead of the period or week on due.

Minimal errors on reporting. Since the BI tool is managed and controlled by a large team of experienced engineers, it’s not easy to mess. Every engineer and analyst does the maker checker. Before one process is taken to the next, the period process checks for errors and alerts the staff working on that process, the process shall not proceed if the current process is not solved. This kind of scenario ensures that the final process only gives quality reporting. On the contrary though, this is not easier to achieve when doing manual reporting.

**Literature review**

In an article written by Cohen (2020), there is need for organisations and corporations to ensure satisfactory data storage mechanisms. A lot of organisations do not store their data and neither do they keep track of it. Since data is the future gold of the inquires,, cloud solutions and scalable sever are supposed to be impended to keep track of whichever changes are coming in and then historically start to analyse how the data trend is performing over the years. This is important especially in fields where the subject matter keeps changing frequently. An example here is with health care industry, Organisations need to assess how the diseases pattern is changing with time and be able to use predictive analytics to simulate future variances.

Rabase et al (2020), producing future data scientist who will be able to implement effective BI systems will require an adequate increments in institutions of higher learning by training more data scientist in this field. Currently the world is still in shortage of adequate engineers who understand the value of the changing data field. New concepts are being introduced into the field and the best that organisations can do is to train more data scientist, our world is going to rely on the data we produce and consume today. Concepts such as machine learning, deep learning are still new in most countries and institutions need to train specialist in these fields.

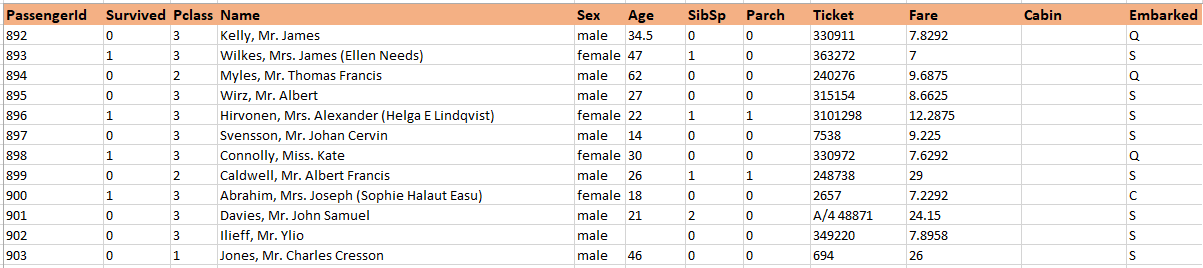
As shared by Lughofer et al (2019), traditional decision support tools are changing and evolving to include a cross organisational 360 degree view. Initially this is the kind of work that was left to CRMS and intelligent systems, but this was only partial. More focus and attention is now being drawn to data modelling and forecasting, this was a big element that was missing in earlier applications of DSS. To achieve this BI elements have had to be incorporate into these legacies.

The earlier literature reviews by these scholars do not wholesomely describe the relationship between big data, legacy systems and their stance in information security, this is still a widely open topic and future researchers can look at.

**Data Analysis and interpretations**

**Dataset and description**

The dataset in use for this particular exercise were those obtained from the kaggle’s <https://www.kaggle.com/brendan45774/test-file>. The dataset is meant for the particular analyses that re required. A quick overview of the dataset is as shown below:



The dataset is a collection of the members of the titanic who were there before it plunged and cased a disaster in 1912.

The variables that are contained in the dataset include the following:

* Passenger ID
* Status of survival
* Passenger class
* Name
* Sex
* Age
* Parch
* Ticket number

**The methodology**

Three tools are going to be used for the data analysis and visualisation:

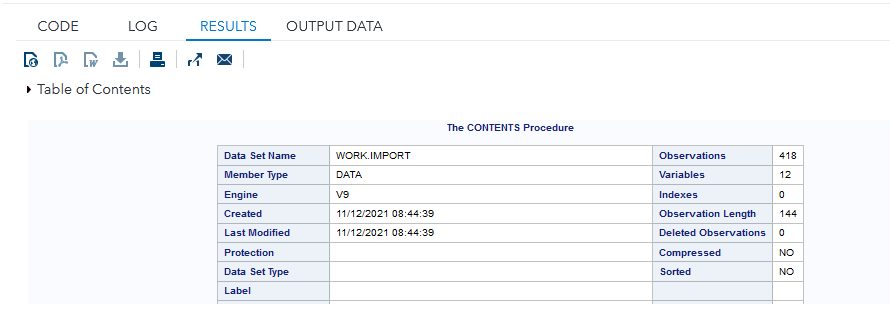
* SAS studio
* Microsoft Excel
* SQL on Dbeaver DBMS

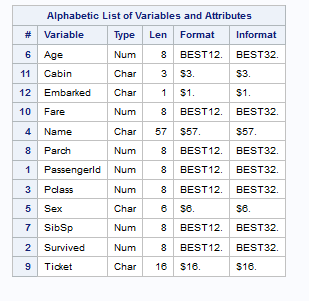
**Aims and objectives**

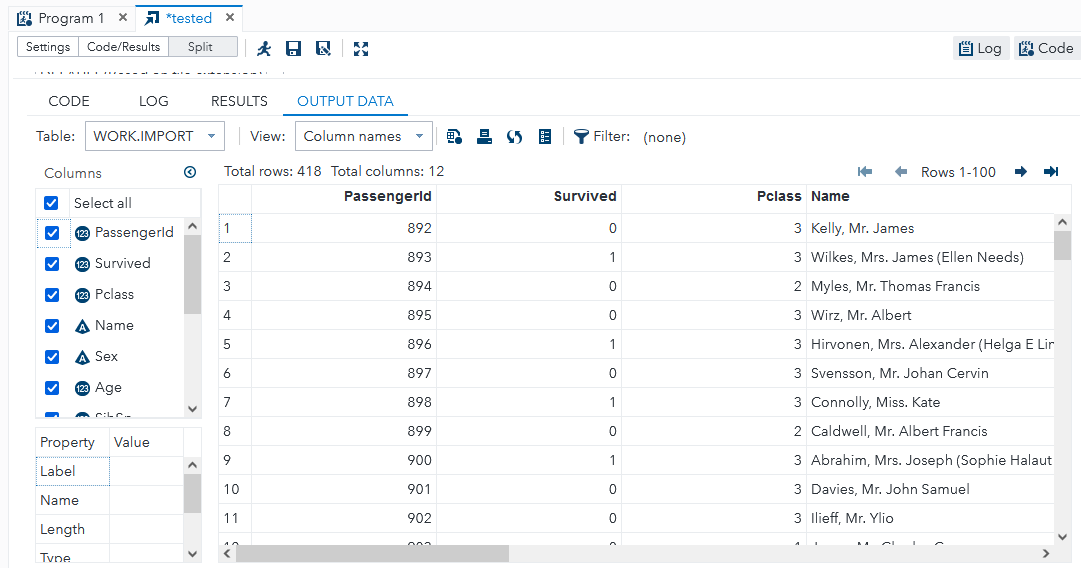
Based on the dataset provided, the following are the aims and objectives of the study:

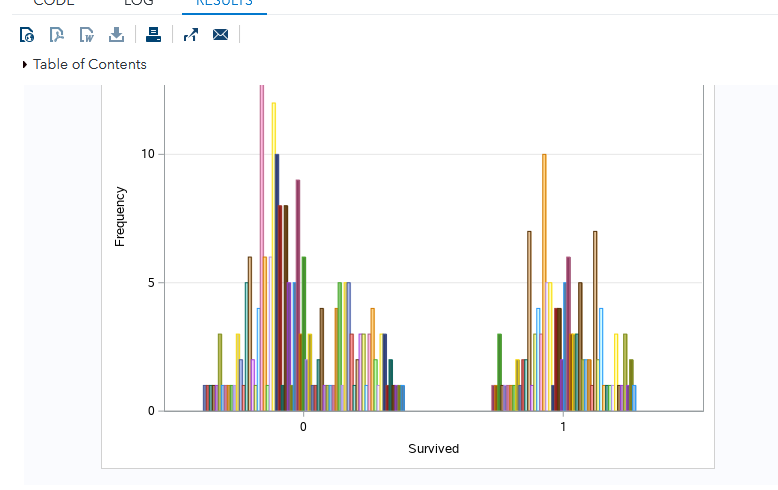
1. To identify a quick statically summary of the dataset using Excel
2. To establish the relationship between the gender and the survival status using SAS studio
3. To visualise the data using Excel
4. To get ad hoc insights and query using the SQL language

**Analysis on SAS studio**

****

****



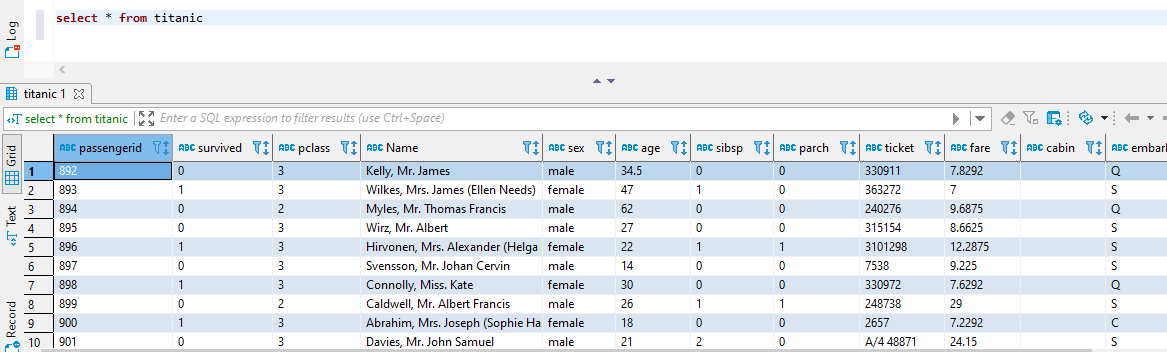


The first step done in the analysis on SAS is to convert the data from CSV and then create its location with the server library. When this is done the software gives the quick summary based on the variables that have been imported. The result is then previewed as table with all the column. The graphical analysis can then show the relationship between those who survived or did not and their age groups.

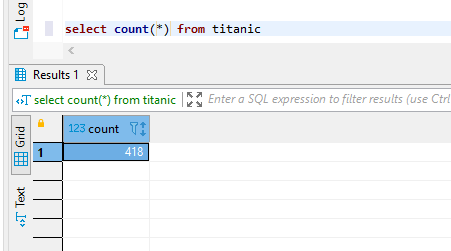
**Data analysis on SQL**

The first step is to upload the dataset on DBMS which in this case is the Dbeaver and then, perform SQL queries on it as below:

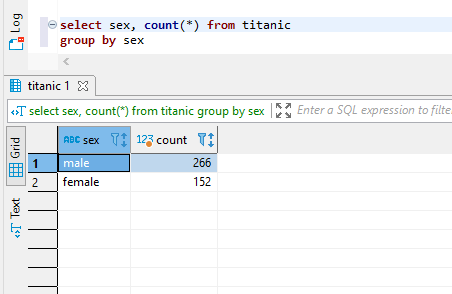
1. Selecting all the records from the table



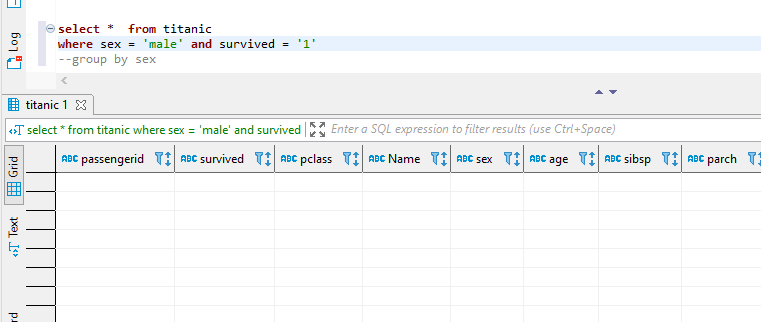
1. Selecting the count of all records from the table;



1. Grouping the passengers by gender

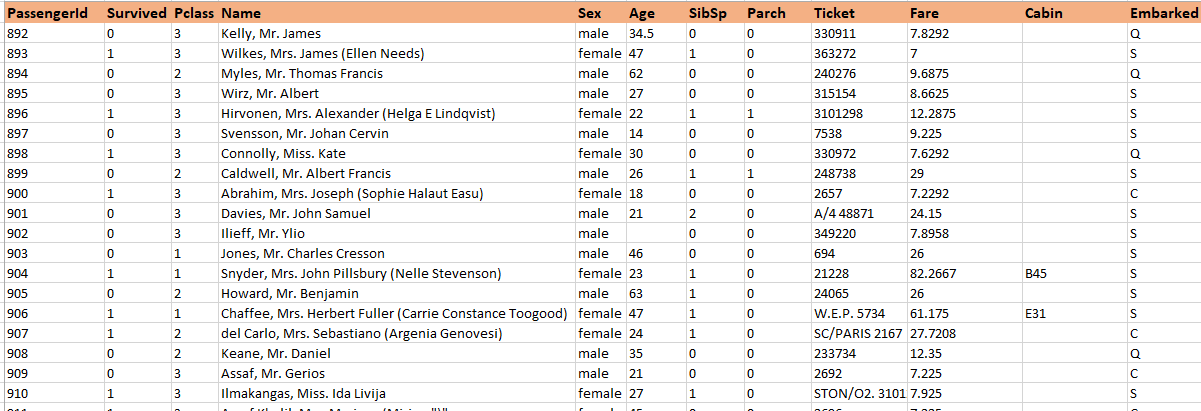


1. Selecting all males who survived ; *no male survived;*

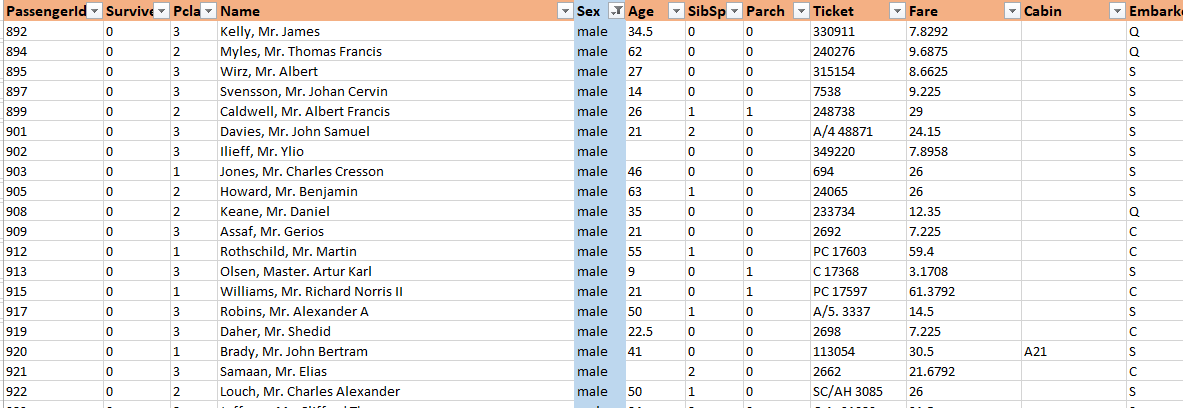


**Data analysis on Excel**

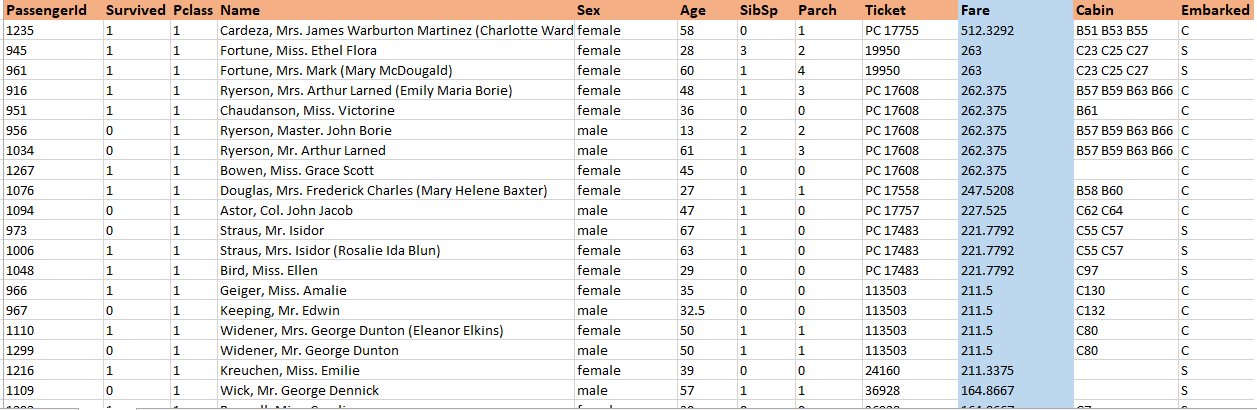
1. Preview on Excel



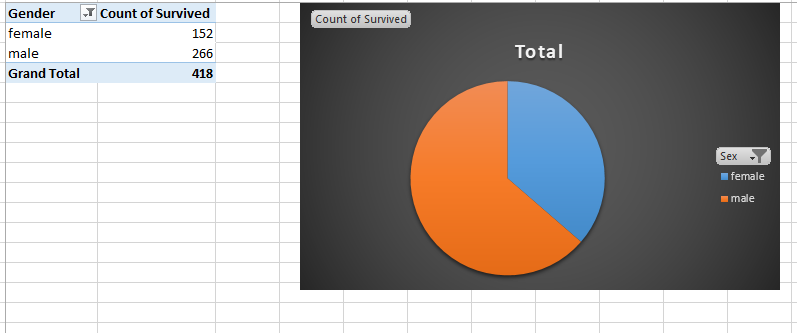
1. Previewing the data based on the gender-male



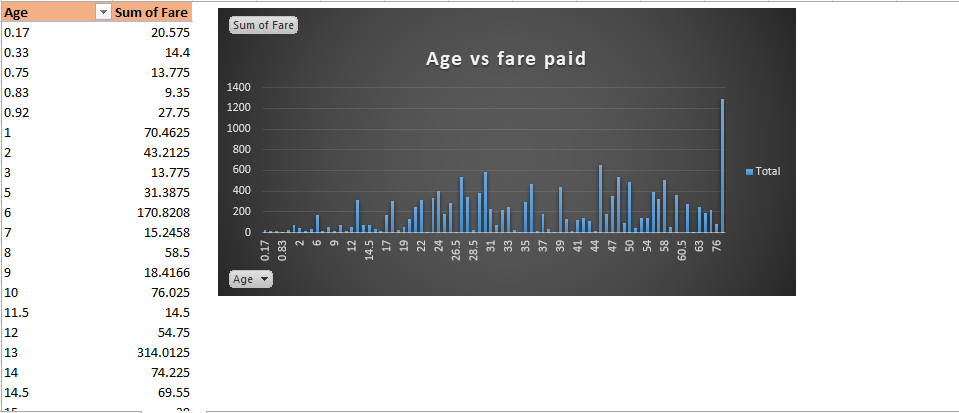
1. Sorting the dataset based on fare amounts paid from the largest to the smallest



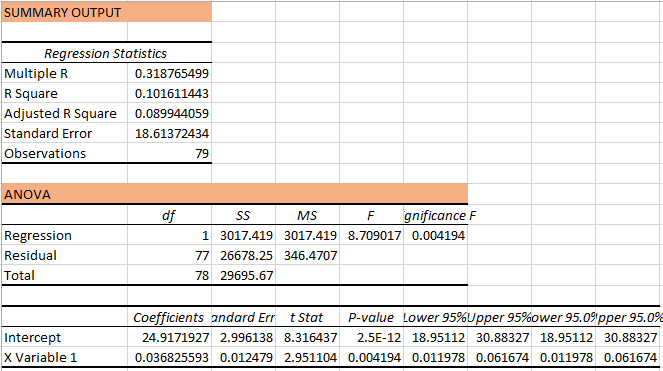
1. Grouping the data via pivots



1. Age vs. fare paid



1. Summary of the dataset. Variables of age and sum of fare paid



The regression level is 1. Meaning that there is a close relationship between the age and the amount of fare paid. If age increased so was the fare.

**Summary and conclusions**

There are quite a number of ways of measuring data. We have quantitative and qualitative data, both have their tools of measurement. R and Python are powerful tools of measuring and predicting data. R is a majorly focused on Statistical analysis. Python is multipurpose. Unstructured database systems is the latest method of data storage even though web 3.0 is till new. Bi applications can be used to leverage company performance based on the strong implementations it has and the various algorithms that can be applied to it. Industries should shift from manual analysis to the adoption of BI system if delivery of output and revenue maximisations are to be considered.

**REFERENCES**

Cohen, R. (2020). Introduction to Data Management, Life History, and Demography.

Corea, F. (2019). Introduction to Data. In An Introduction to Data (pp. 1-5). Springer, Cham.

Rabasa, A., & Heavin, C. (2020). An Introduction to Data Science and Its Applications. In Data Science and Productivity Analytics (pp. 57-81). Springer, Cham.

Cuocolo, R., Perillo, T., De Rosa, E., Ugga, L., & Petretta, M. (2019). Current applications of big data and machine learning in cardiology. Journal of geriatric cardiology: JGC, 16(8), 601.

Lughofer, E., & Sayed-Mouchaweh, M. (Eds.). (2019). Predictive maintenance in dynamic systems: advanced methods, decision support tools and real-world applications. Springer.

Tran, T. O. (2020). Database Systems.

Ilin, I., Klimin, A., & Shaban, A. (2019). Features of Big Data approach and new opportunities of BI-systems in marketing activities. In E3S Web of Conferences (Vol. 110, p. 02054). EDP Sciences.

Chiang, R. H., Grover, V., Liang, T. P., & Zhang, D. (2018). Strategic value of big data and business analytics.

Golding, N. (2019). greta: simple and scalable statistical modelling in R. Journal of Open Source Software, 4(40), 1601.