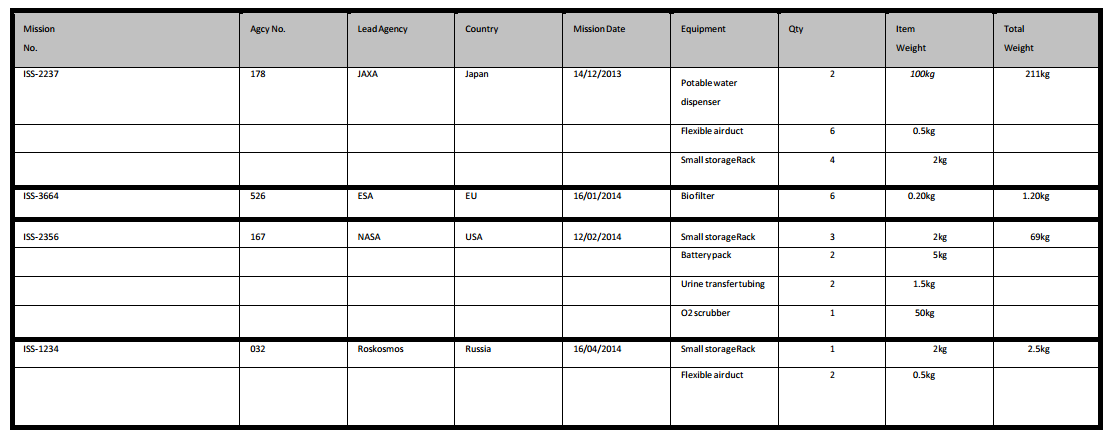
**220CT – Data and Information Retrieval**

**Coursework**

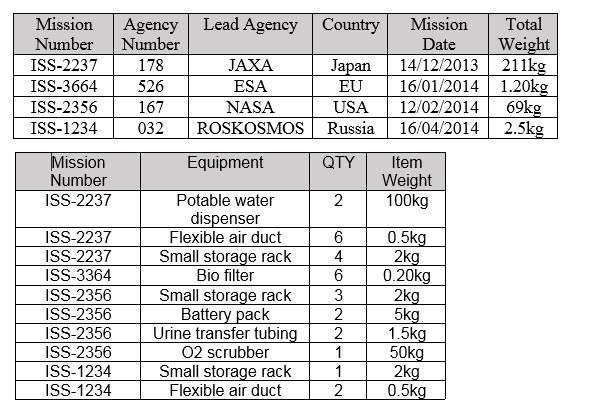
*VENELIN DANIELOV DIMITROV  
ID No: 6297262*

**Task 1, Part 1: DB Design (15 marks)***The International Space Station (ISS) is a habitable artificial satellite in low Earth orbit. It is the ninth space station to be inhabited by crews following previous orbital stations that were launched by the US the former Soviet Union and later Russia. The ISS is intended to be a laboratory, observatory and factory in space as well as to provide transportation, maintenance, and act as a staging base for possible future missions to the Moon, Mars and beyond. In order to support the crew and overall operation of ISS the space agencies in charge of running the station conduct regular missions to launch spacecraft carrying payloads of essential or replacement equipment up to ISS. A payload inventory, see table below, is recorded of each mission, consisting of the space agency leading the mission and the equipment payload to be sent up to ISS. The overall weight of the payload is also determined in order to calculate the fuel needed for orbital insertion of the spacecraft to successfully rendezvous with ISS.*

**

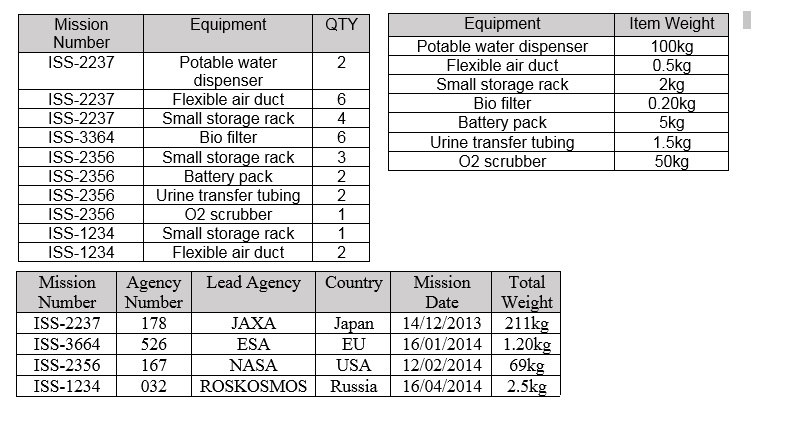
*Deliverable: Using SQL, implement the database above. To do so, you’ll need to normalise the table (should that be required), identify the attributes, create and Entity-Relationship Diagram and create the tables using SQL commands.*  
**SOLUTION:**   
For this graph I would use traditional relation DB form because it’s a small set of data entities and can easily be split to separate tables using normalisation.

**1NF**

The main table isn’t in the first form because “Equipment” had a lot of values stored. So the solution to this one will be to create a new table called “Equipment” or “Inventory” with the only contents (Mission Number, Equipment, Quantity and Item Weight) gathered from the original example. And the other one containing the “Agency” and “Mission related information” as it is the most important information left out from the creation of the first table.  
  


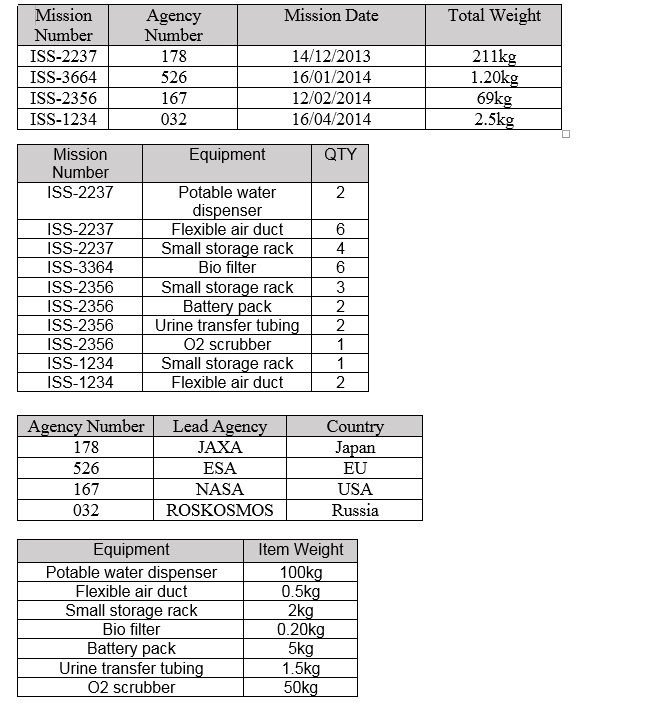
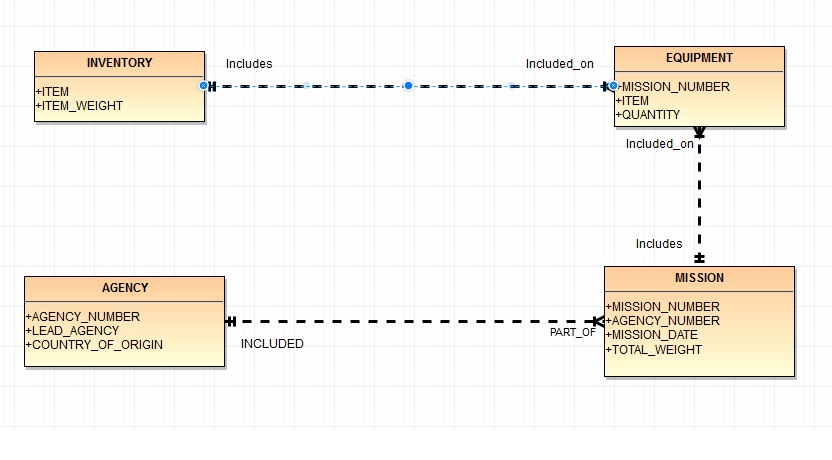
**2NF**

If I want it to be able to transform it from first form (1NF) into second form (2NF), any of the attributes depending only a part of a single table have to be removed and put into a third new one. Therefore the new table will hold the contents of equipment and weight from the “Equipment” table from 1NF. The repeated elements will be erased. The agency table stays at is for this form as there is nothing that is duplicated or needed to be transferred.



**3NF**

In order to transform from 2NF to 3NF, we have to remove any attributes that are more dependent on other non-key attributes and place them in a new table. The new table will be called “Agency”. Looking after that at the table “Equipment” we see that I has no primary key because there are no unique attributes, so the next move will be to put two foreign keys to link the two tables – “Inventory” and “Missions”

  
  
**E-R DiagramThere are **

The four entities in this diagram. Agency, Inventory, Mission and Equipment share an entity relationship with each other as shown in the graph.

* The “Agency” table has to have at least one mission but it is possible if another one appears later so that means it may have many.
* The “Equipment” table has to have at least one “item” from inventory, but it may have more than that depending on the mission.
* “Mission” may have one “equipment” item or depending on the “Mission number” may have more.
* “Agency” table must have at least one “Agency number” included from “Mission” table but it can take multiple requests depending on the “mission date”.

**Implementing in SQL:**

CREATE TABLE mission(

mission\_number VARCHAR2(8) PRIMARY KEY,

agency\_number NUMBER(3) NOT NULL,

mission\_date DATE NOT NULL,

total\_weight NUMBER(5,2) NOT NULL);

CREATE TABLE inventory(

item VARCHAR2(23) PRIMARY KEY,

item\_weight NUMBER(5,2) NOT NULL);

CREATE TABLE equipment(

mission\_number VARCHAR2(8) NOT NULL,

item VARCHAR2(25) NOT NULL,

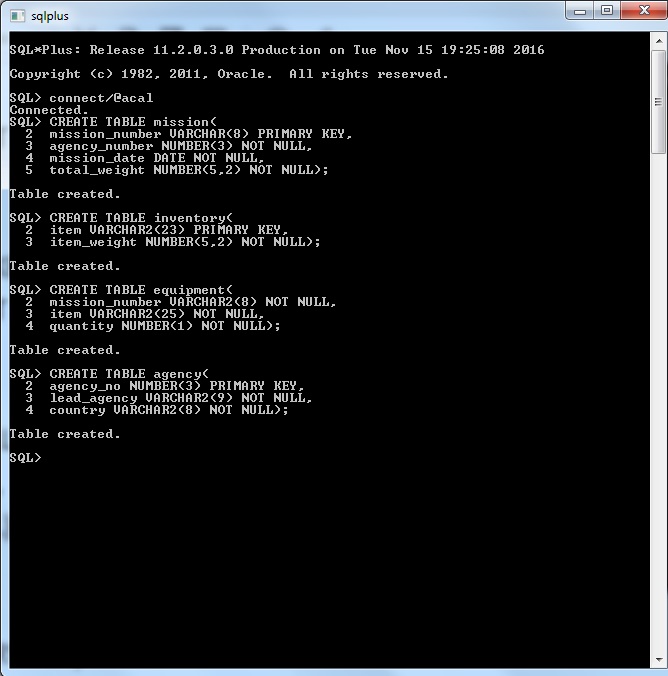
quantity NUMBER(1) NOT NULL);

CREATE TABLE agency(

agency\_number NUMBER(3) PRIMARY KEY,

lead\_agency VARCHAR2(9) NOT NULL,

country VARCHAR2(8) NOT NULL);



ALTER TABLE mission

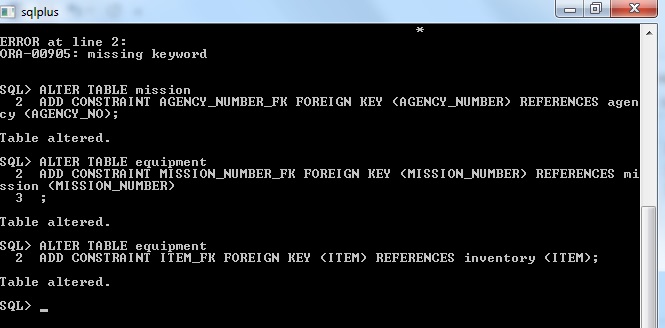
ADD CONSTRAINT AGENCY\_NUMBER\_FK FOREIGN KEY (AGENCY\_NUMBER) REFERENCES agency (AGENCY\_NO);

ALTER TABLE equipment

ADD CONSTRAINT MISSION\_NUMBER\_FK FOREIGN KEY (MIS\_NUMBER) REFERENCES mission (MISSION\_NUMBER);

ALTER TABLE equipment

ADD CONSTRAINT ITEM\_FK FOREIGN KEY (ITEM) REFERENCES inventory (ITEM);

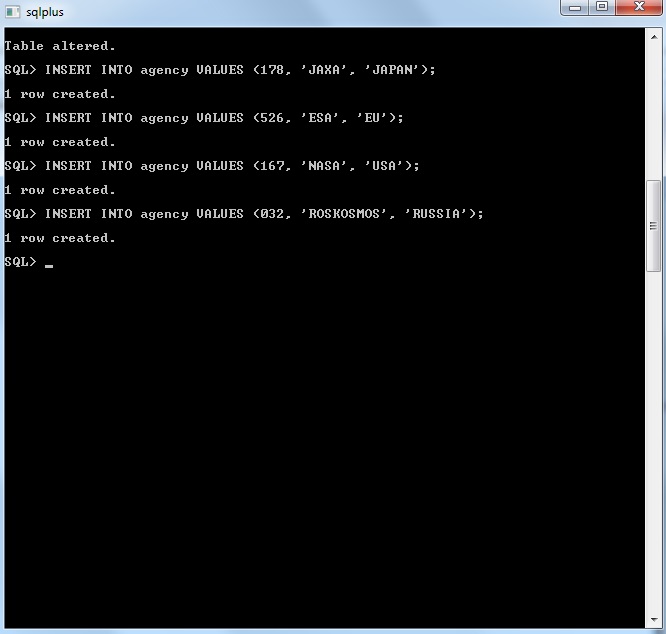
****

INSERT INTO agency VALUES (178, 'JAXA', 'JAPAN');

INSERT INTO agency VALUES (526, 'ESA', 'EU');

INSERT INTO agency VALUES (167, 'NASA', 'USA');

INSERT INTO agency VALUES (032, 'ROSKOSMOS', 'RUSSIA');

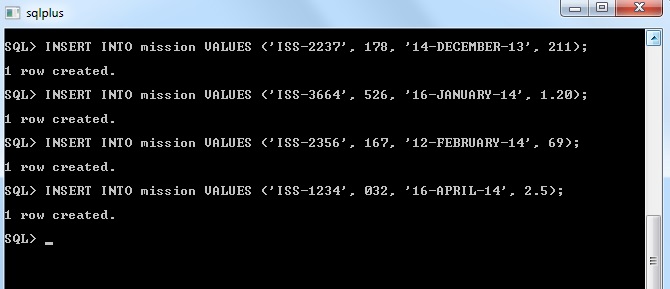
****

INSERT INTO mission VALUES ('ISS-2237', 178, '14-December-13', 211);

INSERT INTO mission VALUES ('ISS-3664', 526, '16-January-14', 1.20);

INSERT INTO mission VALUES ('ISS-2356', 167, '12-February-14', 69);

INSERT INTO mission VALUES ('ISS-1234', 032, '16-April-14', 2.5);

****

ALTER TABLE inventory

MODIFY item VARCHAR2 (30);   
**(\*Saw that the first item didn’t have enough characters assigned so I modified it to accept a number of 30 characters)**  
  
INSERT INTO inventory VALUES ('Portable water dispenser', 100);

INSERT INTO inventory VALUES ('Flexible air duct', 0.5);

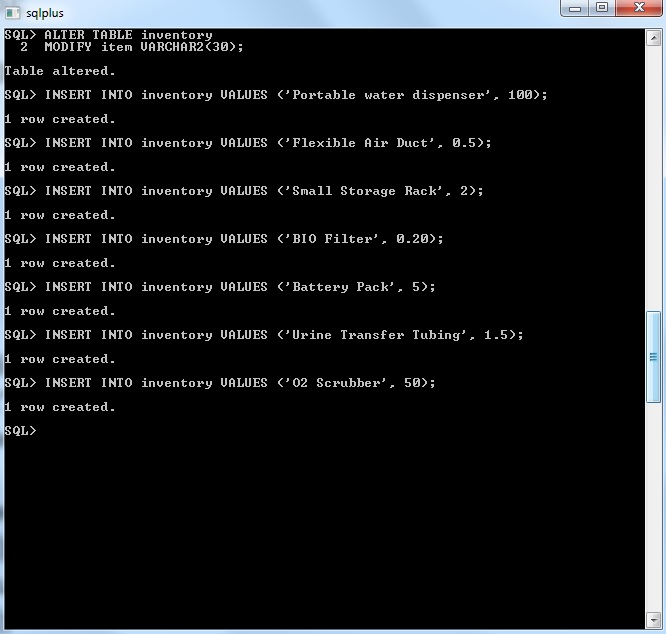
INSERT INTO inventory VALUES ('Small storage rack', 2);

INSERT INTO inventory VALUES ('Bio filter', 0.20);

INSERT INTO inventory VALUES ('Battery Pack', 5);

INSERT INTO inventory VALUES ('Urine transfer tubing', 1.5);

INSERT INTO inventory VALUES ('O2 Scrubber', 50);

****

INSERT INTO equipment VALUES ('ISS-2237', 'Portable water dispenser', 2);

INSERT INTO equipment VALUES ('ISS-2237', 'Flexible air duct', 6);

INSERT INTO equipment VALUES ('ISS-2237', 'Small storage rack', 4);

INSERT INTO equipment VALUES ('ISS-3664', 'Bio filter', 6);

INSERT INTO equipment VALUES ('ISS-2356', 'Small storage rack', 3);

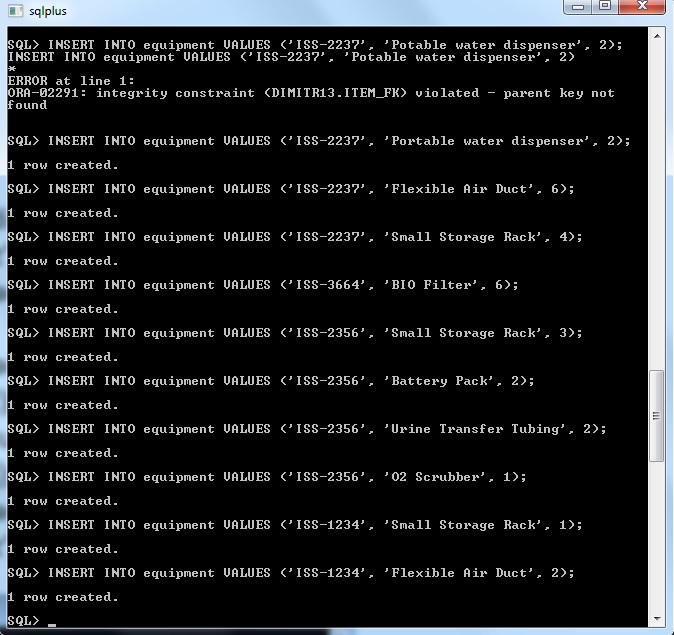
INSERT INTO equipment VALUES ('ISS-2356', 'Battery Pack', 2);

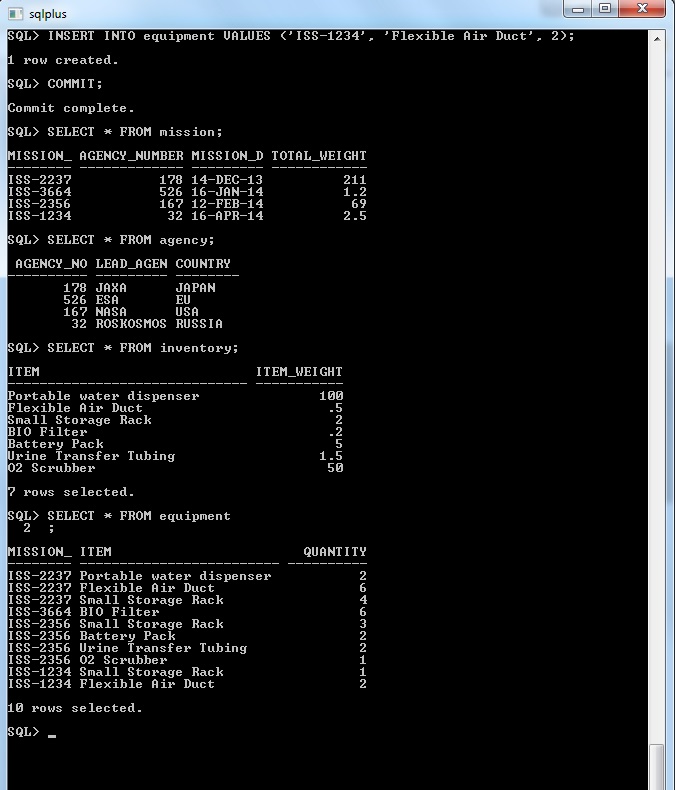
INSERT INTO equipment VALUES ('ISS-2356', 'Urine transfer tubing', 2);

INSERT INTO equipment VALUES ('ISS-2356', 'O2 Scrubber', 1);

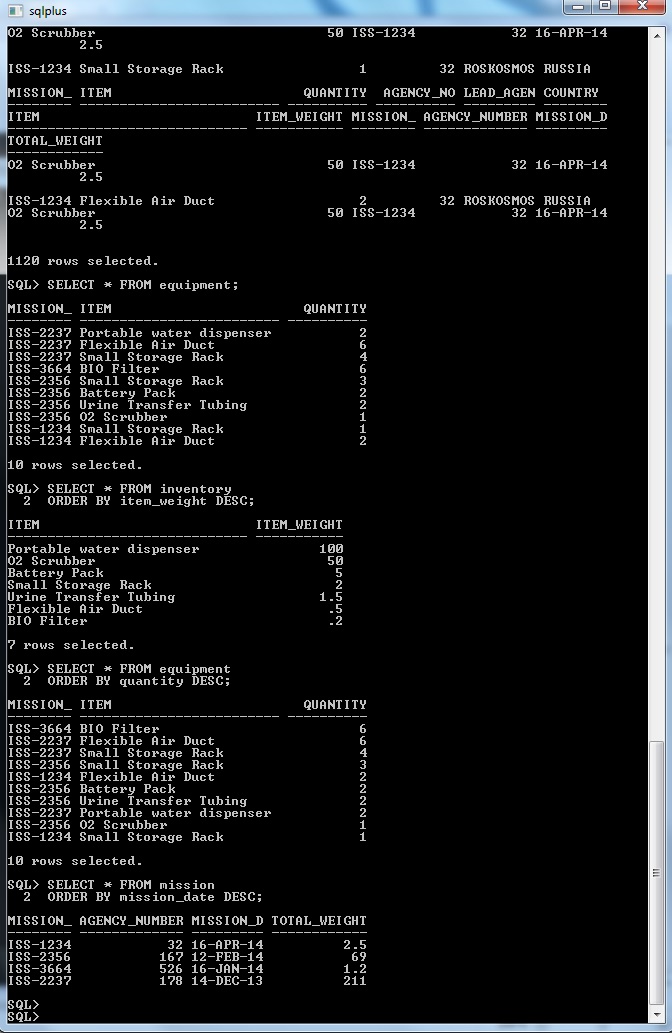
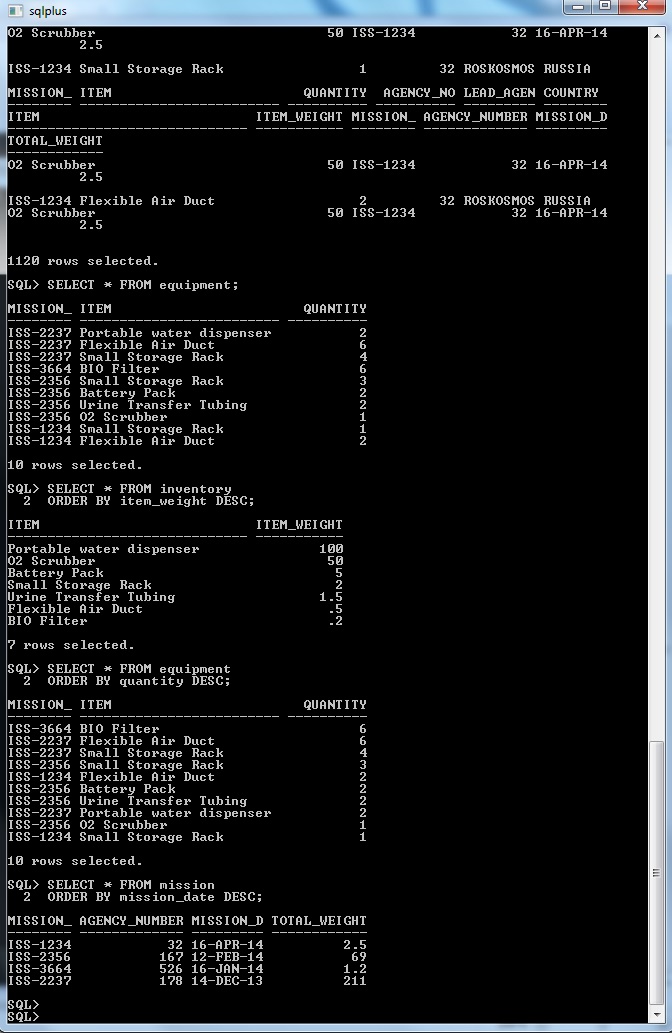
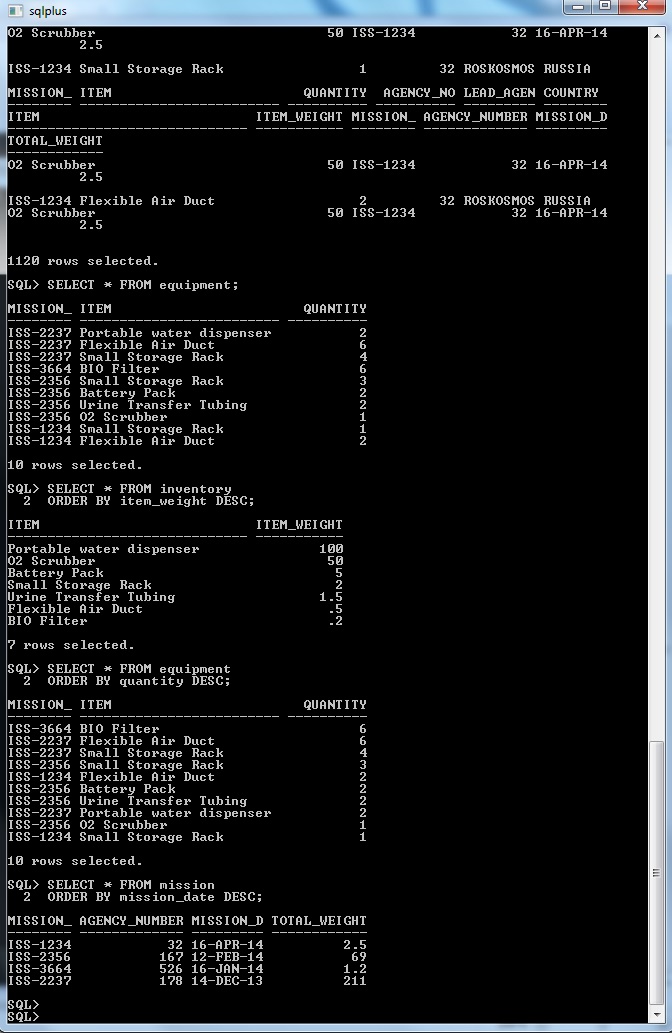
INSERT INTO equipment VALUES ('ISS-1234', 'Small storage rack', 1);

INSERT INTO equipment VALUES ('ISS-1234', 'Flexible air duct', 2);

****

**Commit and final look of the tables.  
 **

**Examples for query with the working tables.**

1. Produce an order of the items from the “inventory” table sorted in descending order of their weight.
2. Produce a list of the items in table “equipment” and sort them by their quantity for the acquired mission in descending order.
3. Produce a list of the items in table “mission”, ordered by their date of execution in descending order.  
   

**Task 1, Part 2:**

The NASA exoplanet dataset archive can be found here:

http://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=planets

You are asked to design a Database solution for the data above. What would you use? How? Why did you make that choice? What are the advantages and disadvantages?

Deliverable:

Your solution must include the following:

1. The DB solution of your choice.

2. A detailed explanation of how these data will be stored and accessed in the DB you suggest.

3. The benefits of this solution in relation to the data above and its size.

4. The QoS (such as scalability) provided/should be provided to the user should this solution be adopted.

For this task I have chosen to use Cassandra DB. The set which needs to be evaluated contains a huge amount of data with numbers calculating different kind of values like mass, axis and orbital days which can go for more than a billion in value. Each row is showing specific data collected for each exoplanet discovered by NASA and the other space agencies. The information about it includes – name of the planet, how it was discovered, orbital period, mass, distance from us, temperature and etc.  
  
The table is being updated daily for the most recent findings which require a real-time analytics database and Cassandra DB as a column oriented database is one of the best at this sort of things. By different given benchmarks it supresses other DBs in calculating big data sets and updating them in real time due to its ability of supporting heavy write operations. Before adding each information gathered to the table it is been calculated and stored in the DB which the program does under JSON and you can either choose to do it in a text file or a blob. And if you want to see any of the report which was saved you can have your query defined accordingly and will generate your data at real time.  
  
One of the big pluses of choosing this DB is its function of integrating with Hadoop and Hive tools, mainly because the program was written under Java. Data will sometimes needs to be applied and calculated then stored in the table which can take more than a few hours and this process can be done while the program is idle and the user has done implementing his sets. While the amounts of the data which are collected are in enormous sizes single events may result in thousands of insertions. All the data is being stored in a data structure located in the memory or in logs and after that is being flushed to a more read-permanent and read-optimized file which can be accessed for a later time. This has to be one of the “claim to fame” for the app as not many can be proud of this function of write speed.   
  
Beneath the covers, the storage layers for Cassandra is just basically a key/value storage system. This means that you will have to organize the input data mostly around the queries that you want to view rather than around the whole structure. It has a limited support for aggregations for a single partition and has an unpredictable performance due to the processes it does in the background. Which means it does not work very well on an existing applications so it’s better used from the starting stages in the early development of the project.

**Task 2: Presentation (20 marks)**

You will be required to conduct research one of the subjects below and explain how big data/data science is being currently used as a solution to help prevent them. Then present your arguments on the ethical and privacy issues you may encounter.

Domains:

* Fraud detection
* Phishing detection
* Identity theft

Deliverable:

A scientific poster containing the following sections:

* A description of the domain/problem and why you chose it.
* A description of how Big Data/Data Science is being used within the domain, including an overall description of the specific techniques and technologies that are used (showing evidence of research).
* Reflection on whether Big Data/Data Science solutions are successfully meeting business objectives.
* A description on how you think that Big Data/Data Science can be further used within that domain.
* Analysis of how the Big Data/Data Science ideas and solutions in your domain of study could be expanded to other Security domains and how knowledge and experience can be transferred.
* Conclusion and closing remarks.

As people living in the information age we are required to use internet daily. That means we are visiting different kinds of websites daily, which for some we are unaware of their origins but we can easily be lured into giving personal information in order to access them and view their content. Many have security protections for users and every detail which was input is stored in a hash key on a server and is unreadable to administrators. But the treat of being scammed is not only in the virtual space and in fact the graph below shows that it is more common your security to be breached   
  
There are pages which collect data, information and even sometimes personal records and have the power to sell it to other media or government authorities for personal gain. This is also possible of happening through personal mail, shopping discount cards or street surveys. In those sheets you are required to input your names, address, age, date of birth which are then copied down to servers of the company doing the research. The big data is then sold or even in some cases given for free to other companies for research or advertisement.   
  
There are numerous cases in the US where fraudsters only having the “Social Security Number” of a person can candidate for a loan from a bank and get approved. It was even possible for deceased people to go into debt but since then the government acquired an SSN randomization and stopped sharing the big data daily on their website which was public for everyone to see and mostly used by criminals and hackers and not by many users. Scammers can get the rest of the data through websites collecting that info from companies for small amount of payment, or phone calls pretending to be the victim to different places he was subscribed. Other methods are:

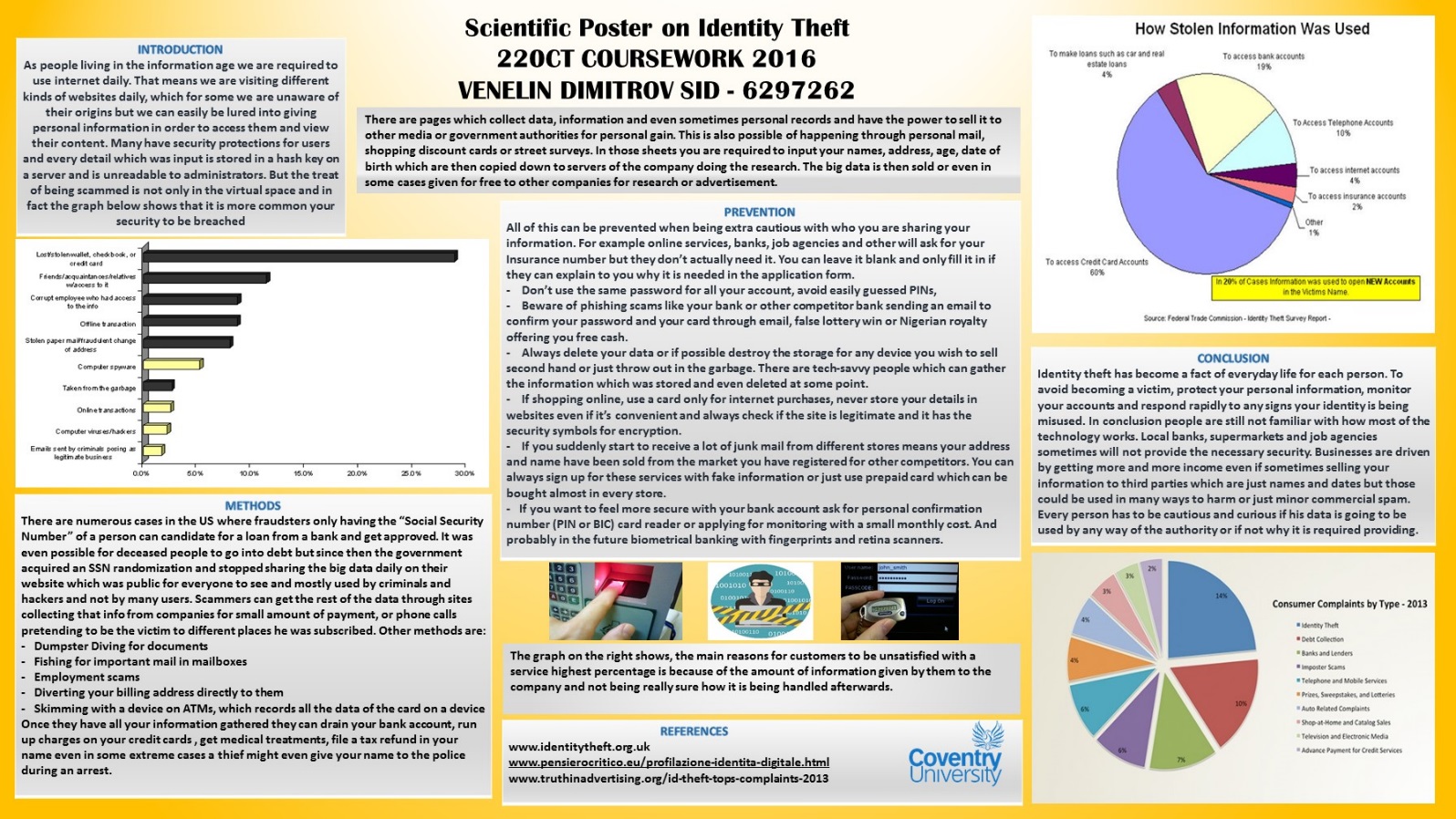
* Dumpster Diving for documents
* Fishing for important mail in mailboxes
* Employment scams
* Diverting your billing address directly to them
* Skimming with a device on ATM machines, which records all the data of the card on a device

Once they have all your information gathered they can drain your bank account, run up charges on your credit cards , get medical treatments, file a tax refund in your name even in some extreme cases a thief might even give your name to the police during an arrest.

All of this can be prevented when being extra cautious with who you are sharing your information. For example online services, banks, job agencies and other will ask for your Insurance number but they don’t actually need it. You can leave it blank and only fill it in if they can explain to you why it is needed in the application form.

* Don’t use the same password for all your account, avoid easily guessed PINs,
* Beware of phishing scams like your bank or other competitor bank sending an email to confirm your password and your card through email, false lottery win or Nigerian royalty offering you free cash.
* Always delete your data or if possible destroy the storage for any device you wish to sell second hand or just throw out in the garbage. There are tech-savvy people which can gather the information which was stored and even deleted at some point.
* If shopping online, use a card only for internet purchases, never store your details in websites even if it’s convenient and always check if the site is legitimate and it has the security symbols for encryption.
* If you suddenly start to receive a lot of junk mail from different stores means your address and name have been sold from the market you have registered for other competitors. You can always sign up for these services with fake information or just use prepaid card which can be bought almost in every store.
  + If you want to feel more secure with your bank account ask for personal confirmation number (PIN or BIC) card reader or applying for monitoring with a small monthly cost. And probably in the future biometrical banking with fingerprints and retina scanners.  
      
    As seen on the top graph main reasons for customers to not be satisfied with a service is because of too amount of information given by them to the company and being really sure how it is being handled afterwards.

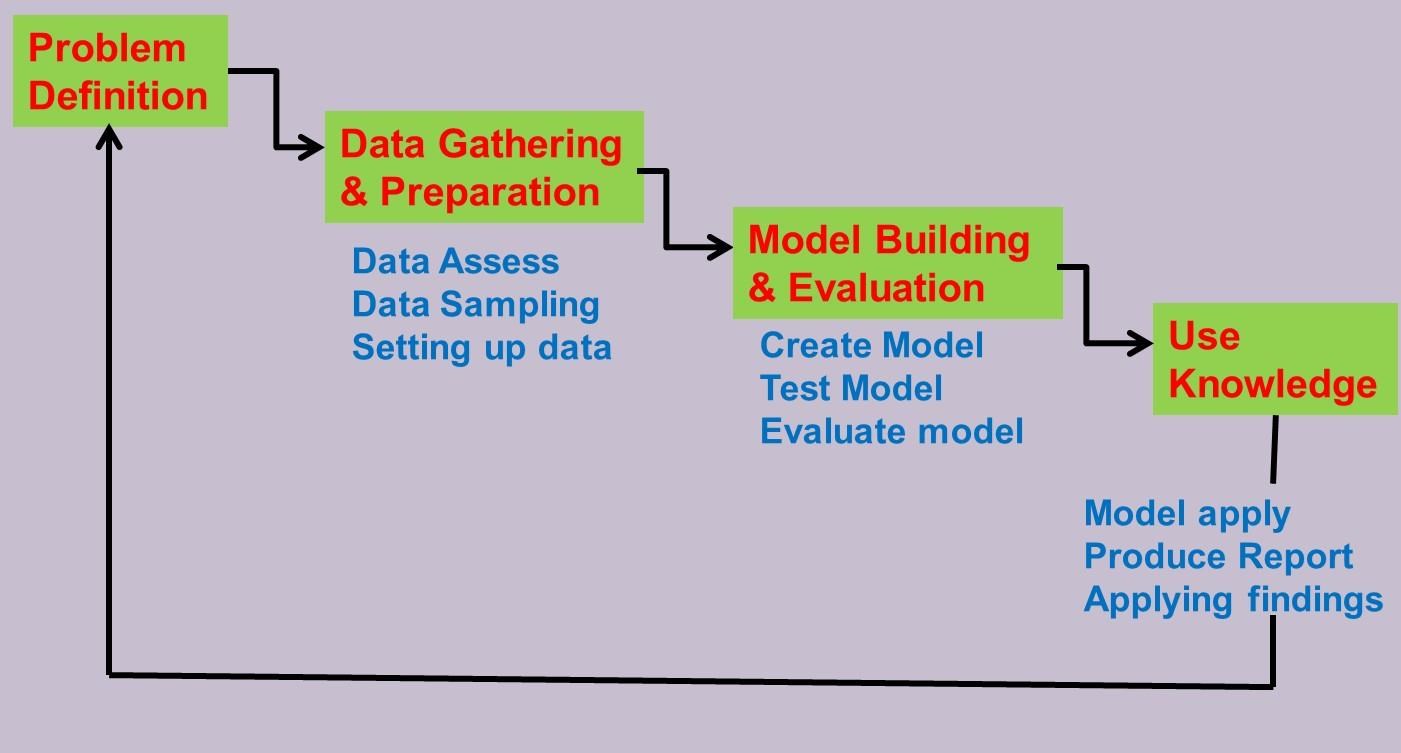
Identity theft has become a fact of everyday life for each person. To avoid becoming a victim, protect your personal information, monitor your accounts and respond rapidly to any signs your identity is being misused. In conclusion people are still not familiar with how most of the technology works. Local banks, supermarkets and job agencies sometimes will not provide the necessary security. Businesses are driven by getting more and more income even if sometimes selling your information to third parties which are just names and dates but those could be used in many ways to harm or just minor commercial spam. Every person has to be cautious and curious if his data is going to be used by any way of the authority or if not why it is required providing.

  
\*The file is included in the submission form.  
  
  
  
**Task 3 – A data mining system for a Hospital (25 marks)**

A hospital has been collecting a great deal of data on their patients and have heard that use of data mining could improve their service. They would like you to create a brief report that includes the following.

(A) What data mining is and an appropriate application for the hospital.   
   
Data mining is used by many companies to calculate big data resources which can discover patterns to make the statistics into something understandable and if understood properly to generate new business ideas, increase revenue, cut the coast or all of them. This would be highly beneficial for the hospital because it could use data mining to detect fraud, access patient’s files, observe high-risk patients and chronic diseases and quarantine after epidemics. It can also be used to find cure for any condition by comparing symptoms, treatments, causes and effects after using certain medication and then analyse which action would be more helpful.

(B) How you would go about creating the system using the data mining lifecycle below.



**Problem Definition**The aim is to maintain loyalty, advertise specific medications and increase number of patients for each practician. This will happen by determining which products should be advertised to specific customers by looking on data how each one has effected on patients with a similar condition on this medication. **Data Gathering & Preparation**The data collected should include the following attributes:

* Vaccination History
* Personal Information
* Current Medical Status
* Insurance Details
* Who to contact in emergencies
* Correspondence about earlier visits and previous GP files
* Allergies

From this information the hospital can gather and make a file with all the important information of the patient necessary for a small or major treatments and can be used by any other authority or clinic which will require this file. It will be gathered by making a short interview with the patient and fulfil all the gathered data from a survey sheet and then transferred to the online database.  **Model Building & Evaluation**

Data mining is required to analyse patterns in the patients dossier, check what medication has been recently used, does he have the required insurance and etc. For example applicants which are over 35 years old are expected to be taking different medications due pain, insomnia or severe cold or could be treated using some anti-biotics. Therefore this data can be used to advertise a drug which is much more effective and does less damage to the system or some new different breed which was not released publicly and follow its development with the current condition.   
  
For instance it is possible to create a model using clusters to determine a brand of medication could be targeted at certain group of people like old people or pensioners. Data-mining could be used to show which types of drug are effective in certain type of blood group or how does it work with different kind of medication and thus keeping the customer safe and not leaving his life in danger for the project relying on recent data.   
  
It could be also used to follow why different patients will choose and go to a different clinic and they could change their GP hours, accessibility or public informing to attract new people signing in. Forecasting if a candidate is going to leave the hospital is also possible by looking at the customers visits, how frequent are they and if he is coming for his check-ups regularly and how much time has he waited outside of the cabinet from his appointed time until the time he was examined.

**Use Knowledge**

With the acquired results, a massive report will be produced by the data miners which will outline the findings of the model. This can be then used to increase the number of patients, make the clinic more famous and examine what most newcomers are after and make those services available and improve them.

(C) If the small amount of data (diabetes.arff) collected so far by the hospital is appropriate for assessing if a person has diabetes.  
  
The patients are evaluated depending on their recent medical survey they have done after testing positive or negative for diabetes. The other information gather for each individual is crucial to finding out what age, sex and blood sugar has to do with the disease and how it can be prevented.  
  
All the subjects in the list are females, with at least 21 years of age from PIMA Indian heritage  
  
1. Number of times pregnant

2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test

3. Diastolic blood pressure (mm Hg)

4. Triceps skin fold thickness (mm)

5. 2-Hour serum insulin (mu U/ml)

6. Body mass index (weight in kg/ (height in m) \*2)

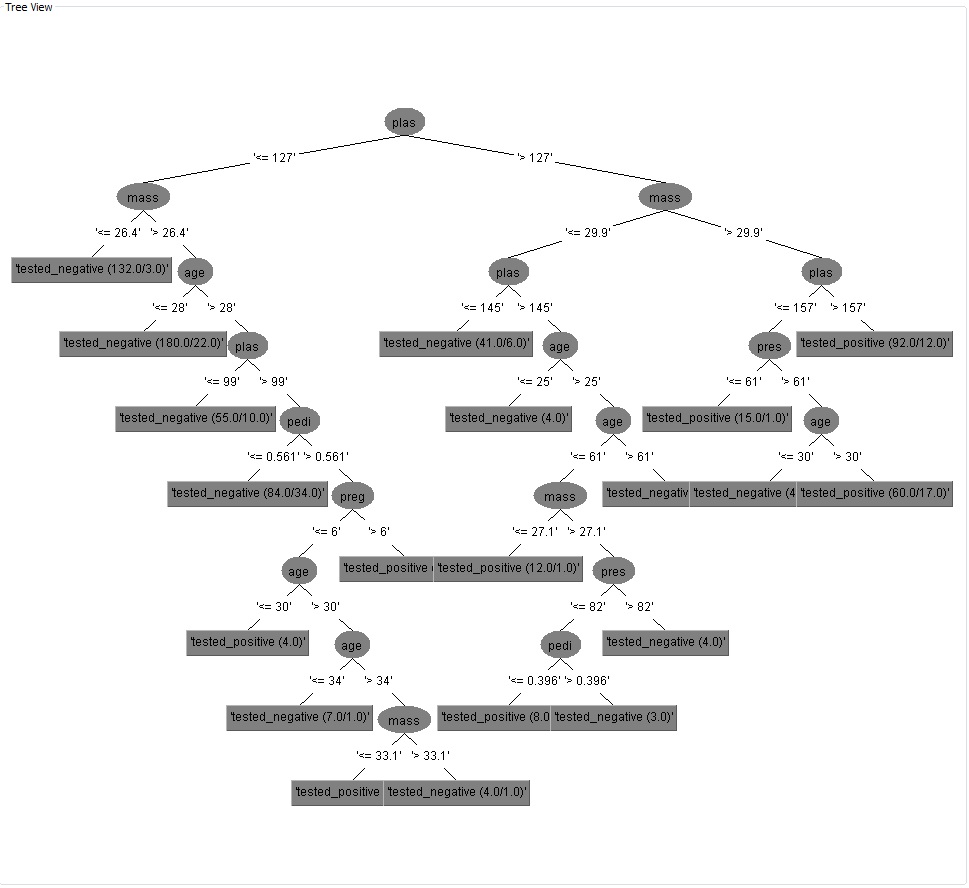
7. Diabetes pedigree function

8. Age (years)

9. Class variable (0 or 1) for negative or positive in diabetes.  
  
These are all attributes and statistics gathered for the study analysis by examine the results from them we can see where the diabetes has more chance to happen. Blood pressure, plasma glucose and the insulin serum can show which type of blood reacts to the results at different ways. Skin thickness, body mass index and times pregnant can gives us results for anything concerning the eternal body mass, muscles and body changes. And diabetes pedigree and age can show how it has advanced depending on age and advancement of the diabetes if present.

(D) The use of a data mining model such as a multilayer perceptron or decision tree to determine whether a person has diabetes. Note, you will need to use a data mining tool like WEKA to create your model and use the diabetes.arff data to train and test this model.

**Decision Tree with all Attributes**



**Summary**

Correctly Classified Instances 567 73.8281 %

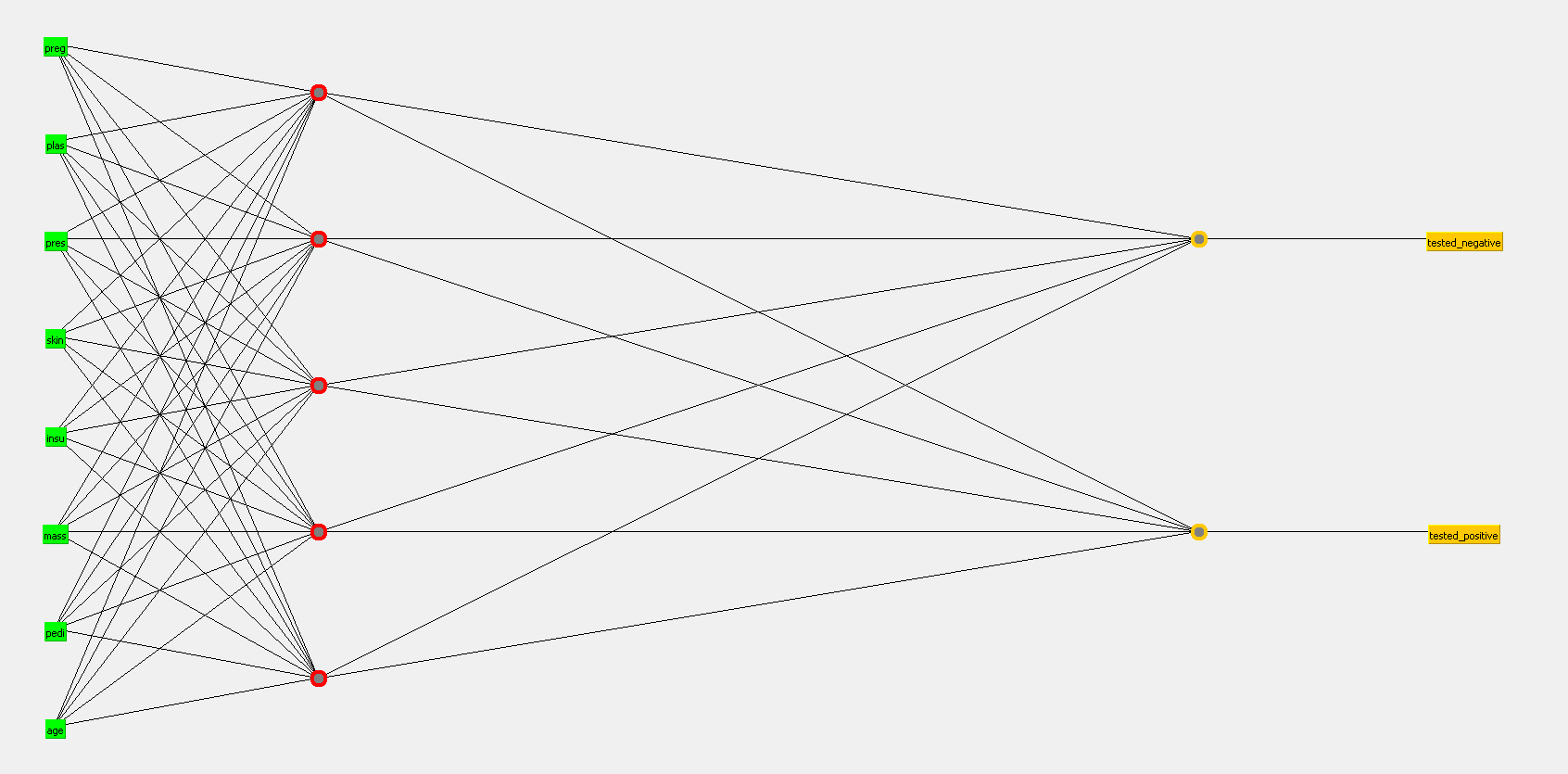
Incorrectly Classified Instances 201 26.1719 %  
Number of total Instances = 768

. === Confusion Matrix ===.

a | b | <-- classified as

407 | 93 | a = Tested Negative

108 |160 | b = Tested Positive  
  
Based on this result 515 are tested “Negative” and 253 had the result as Positive.



Decision Tree when Age is put into five different categories.



=== Summary ===

Correctly Classified Instances 567 73.8281 %

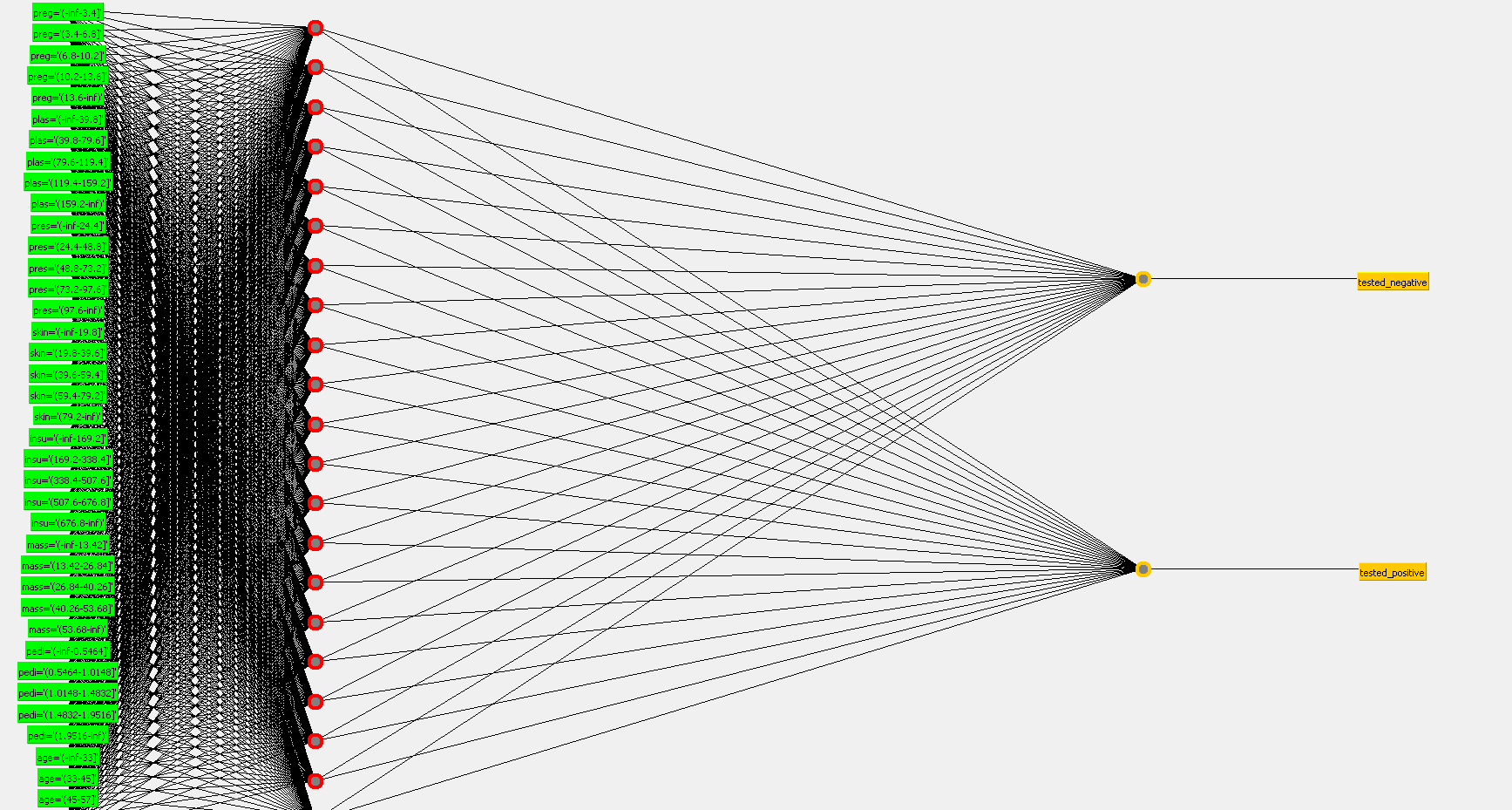
Incorrectly Classified Instances 201 26.1719 %

Total Number of Instances 768

=== Confusion Matrix ===

a b <-- classified as

428 72 | a = tested\_negative

129 139 | b = tested\_positive  
Based on this result 547 are tested “Negative” and 211 had the result as Positive.  
  
  
  
=== Summary ===

Correctly Classified Instances 541 73.8281 %

Incorrectly Classified Instances 227 26.1719 %

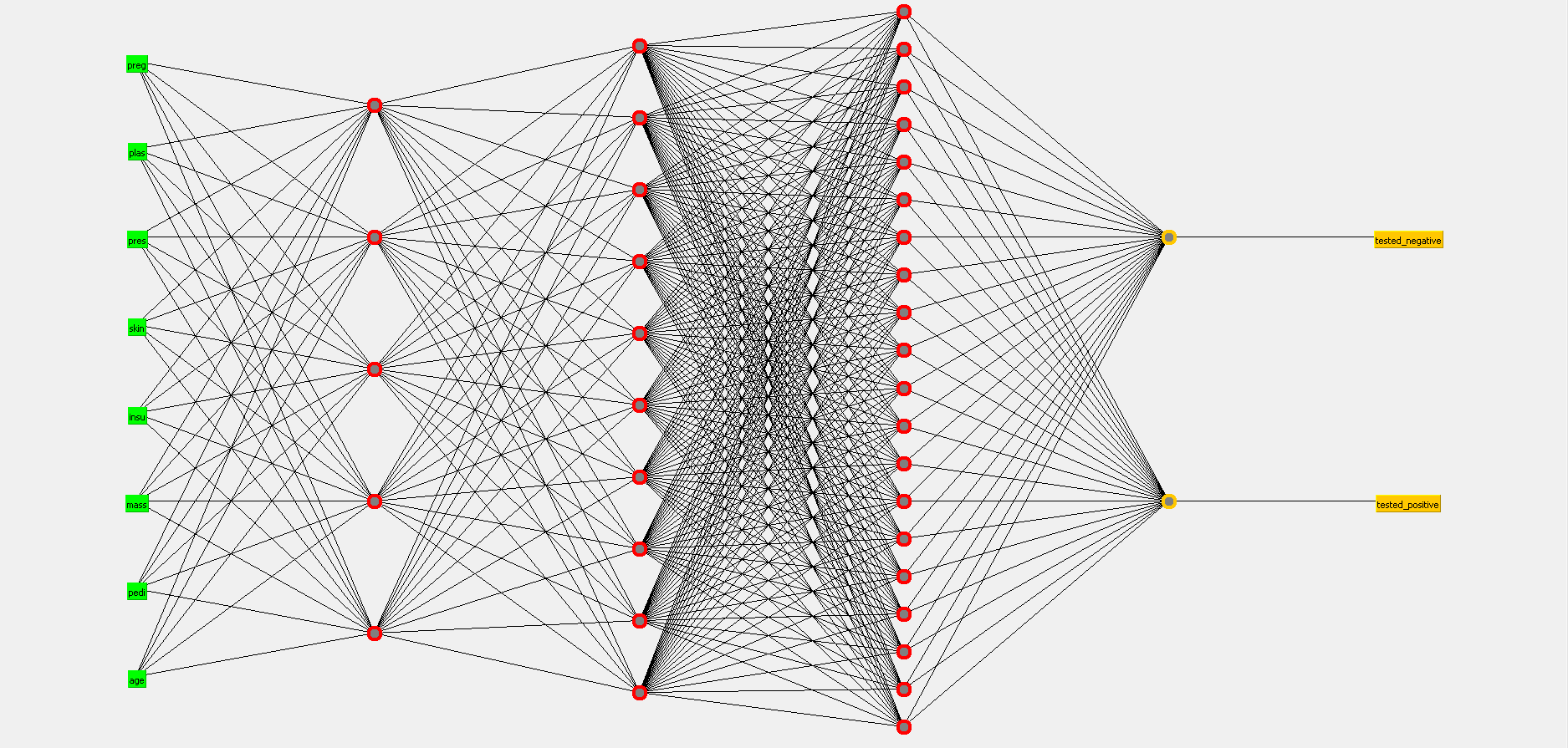
Total Number of Instances 768

=== Confusion Matrix ===

a b <-- classified as

371 95 | a = tested\_negative

151 151 | b = tested\_positive  
  
Based on this result 522 are tested “Negative” and 246 had the result as Positive.  
  
**Hidden Layers Shown**



=== Evaluation on test split ===

=== Summary ===

Correctly Classified Instances 362 70.2913 %

Incorrectly Classified Instances 153 29.7087 %

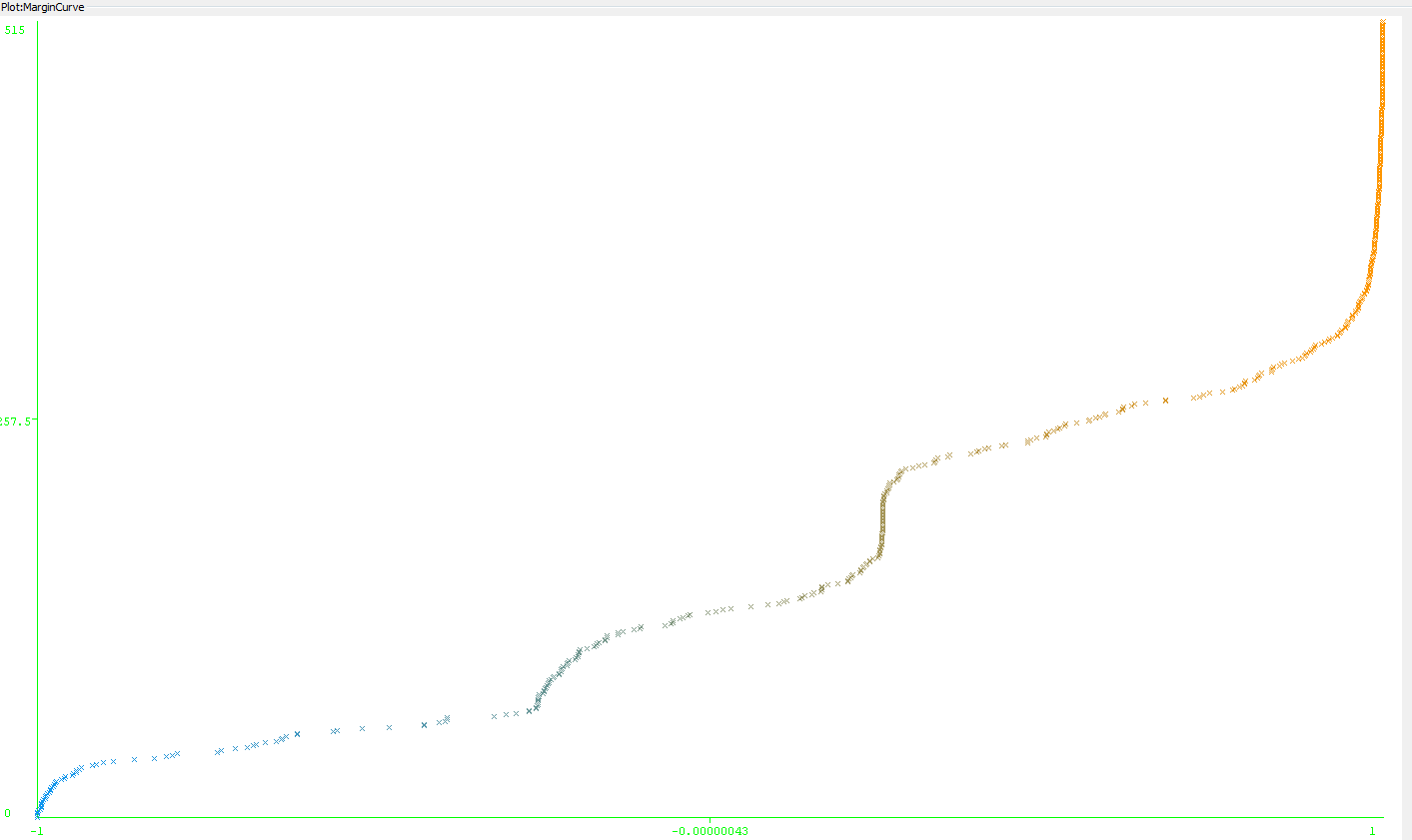
Total Number of Instances 515

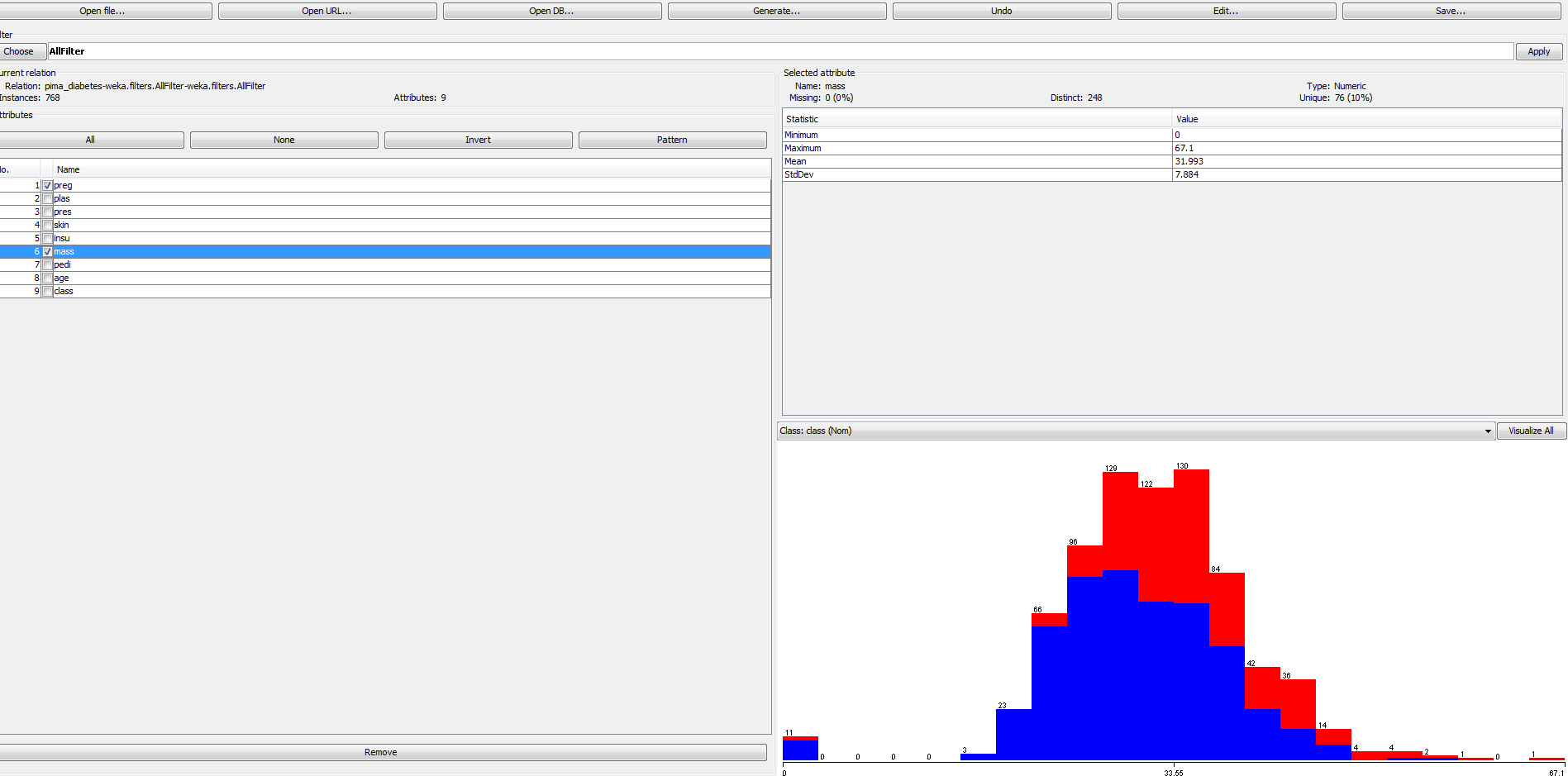
=== Confusion Matrix ===

a b <-- classified as

277 65 | a = tested\_negative

88 85 | b = tested\_positive



  
  
**Conclusion**  
I choose to represent the data with Decision Tree J48 and with Multilayer Perceptron so I would compare all of the models. A decision tree shows the admitted data as a set of decisions leading to different results calculating the data. The Multilayer tree produces a neural network that learns from training and making different computations. All the tasks ended with above 70% accuracy and showed very interesting statistics for each given task.

Looking at the graphs it looks like there is a pattern between BMI, number of pregnancies, pedigree function, and the test results for diabetes. The average BMI did not change as the number of pregnancies increased. Overall those who tested positive for diabetes had higher BMIs than those who did not. There was a lot of empty data containing 0’s which was either for BMI or for ages and that data was isolated from the research as It was interfering with results and not showing correct data.

**Task 4: Your Big Data, Big Idea (25 marks)**

**Aim and Objectives**

**Aim**

The aim of my project is to explore a huge amount of data found on a public website with more than 5000 movie titles with all the information gathered for them through a website called IMDB. The movies can vary from “black and white to color” or from year 1960 to 2016. Also included are all the Facebook page likes of the director and main 3 actor of the film and their names. The possibilities here endless, and it is possible to find any kind of valuable information.

**Objectives**

* Collect the data from the “Internet Movie Data Base - IMDB” (as the data being more than 900gb I decided to find a smaller sample)
* Create and view the data through excel
* Analyse the data
* Visualize the data through – Excell , Data fusions and Google Maps
* Results

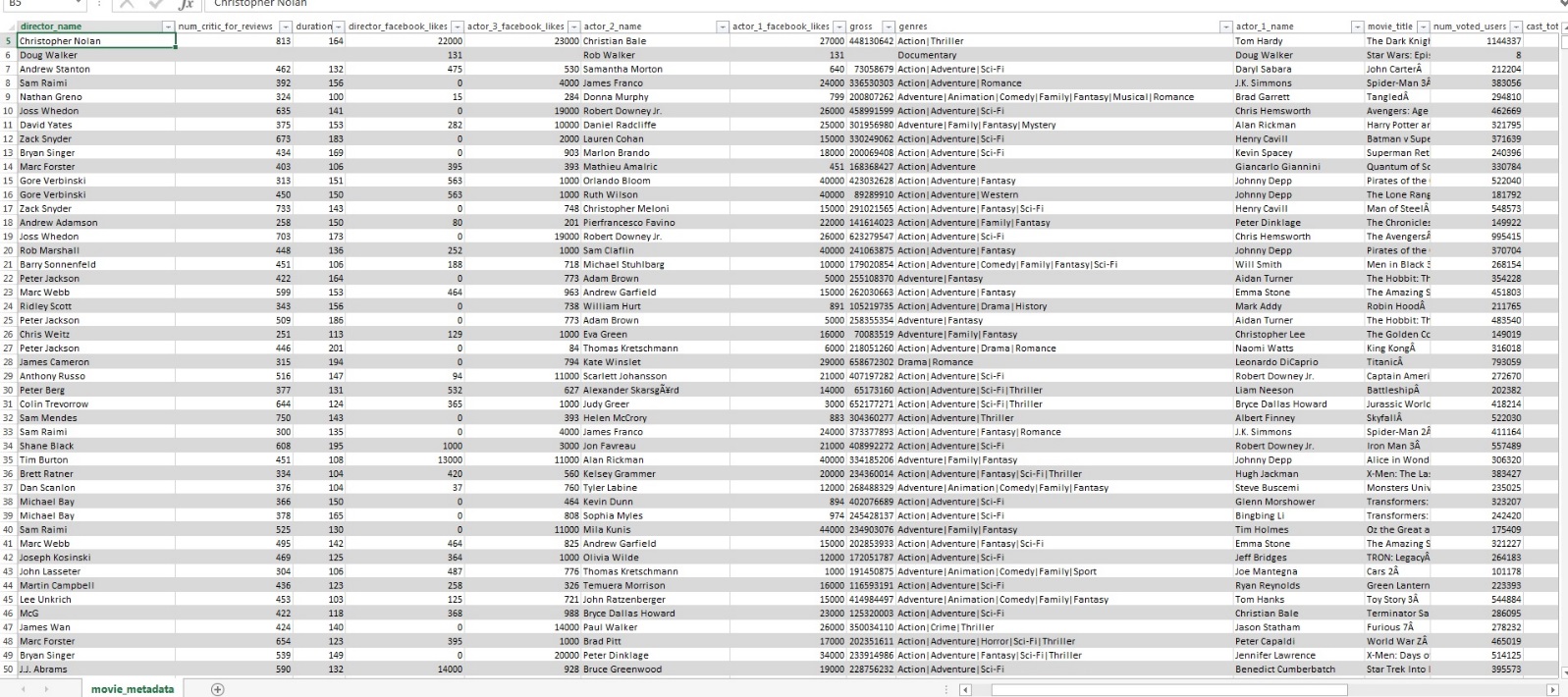
Future developments and ideas  
 **Background**

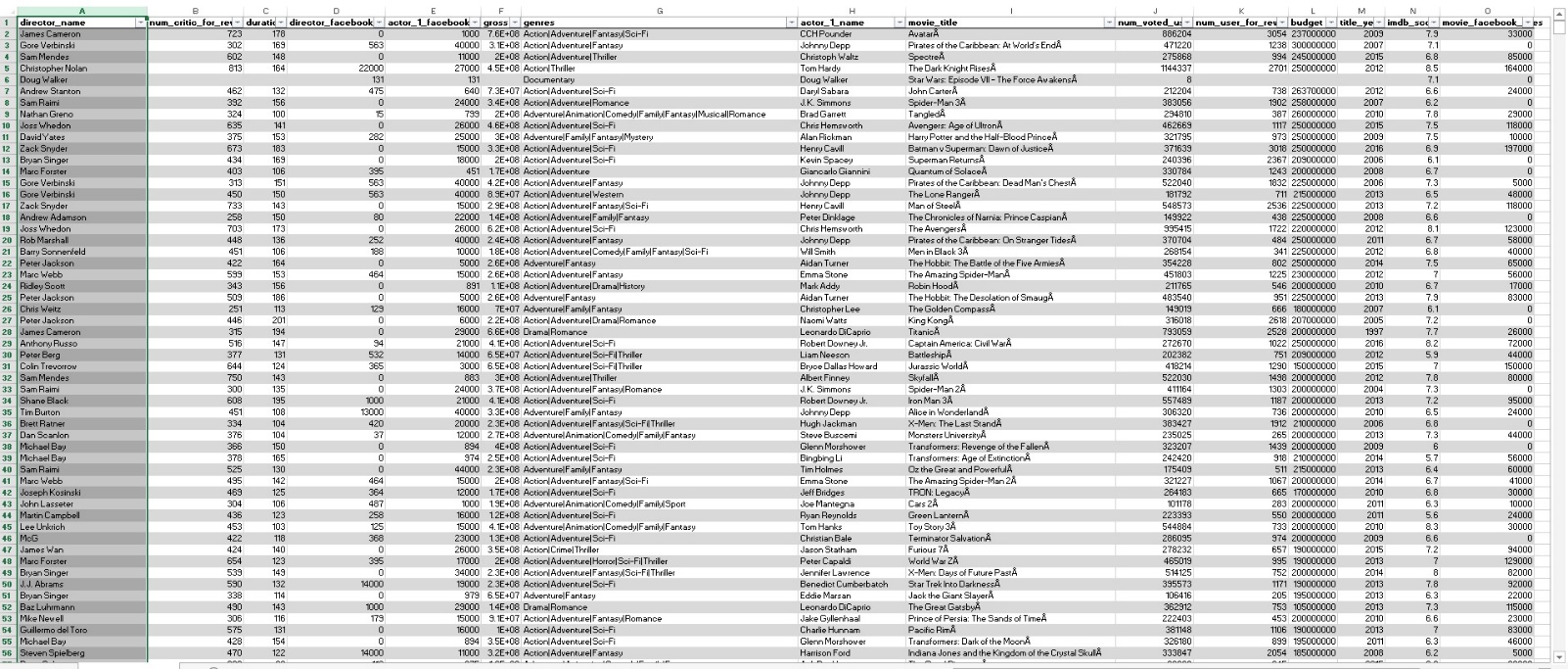
The creation of the movies started in the early 19th century when the motion picture cameras were invented and film production companies started to establish. One of the first movies date to year 1870s, back then the technology had only the power to record black and white pictures with around 100 shots which are then rotated inside a rotor fast enough so it’s shown as a moving character and scenery. As the time progresses many new techniques and ideas became to evolve and the small project movies evolved into a huge enough industry to make other companies and people involved with it. Every product that was made or was in progress was anticipated by the public and had overfilled crowds waiting to enter. Nowadays the technology has evolved that there are new movies coming out daily by companies all around the world, and all of them are with different idea, budget, language, location and genre.  
 **Reasons why I picked this project**

As many other people I am really interested to see the new movie hits and stay with the trends of what is famous and favorited in the public. Also there is a lot of old movies which I am a really big fan of and was interested to see the results of them related to the statistic provided by the big data.

**Acquiring the data**

As mentioned a few paragraphs back I was interested to acquire the whole imdb database but my personal machine was not able to compute those enormous values (and the compressed file was expensive). So I found a 5000 movies sample of the project on a website called Kaggle which has all kinds of different data sets for any type of data which can go from movies, pc games, forum comments, weather for certain cities and etc. As soon as I was able to open the databases in excel I started playing around and making it more simplified to work with.

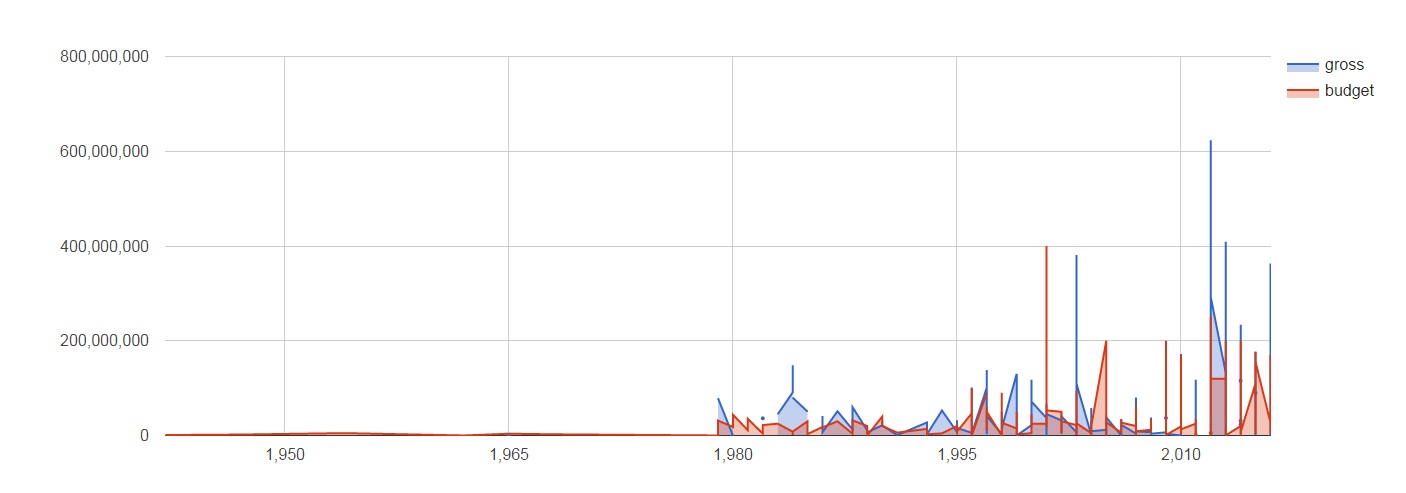


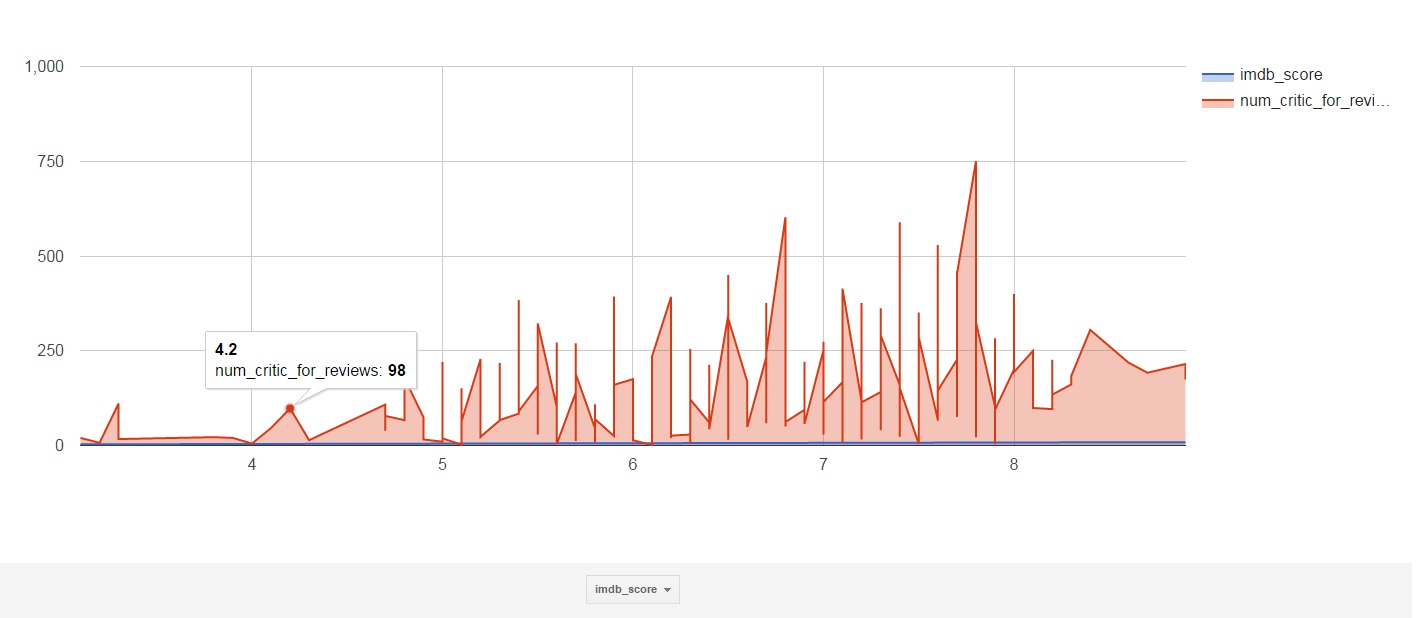
As there was a few tables with not necessary information I decided to delete their contents and make the graph more width which was easy to work with.   
  
Stuff as links for the movie, second, third and fourth actor and their facebook likes has been removed

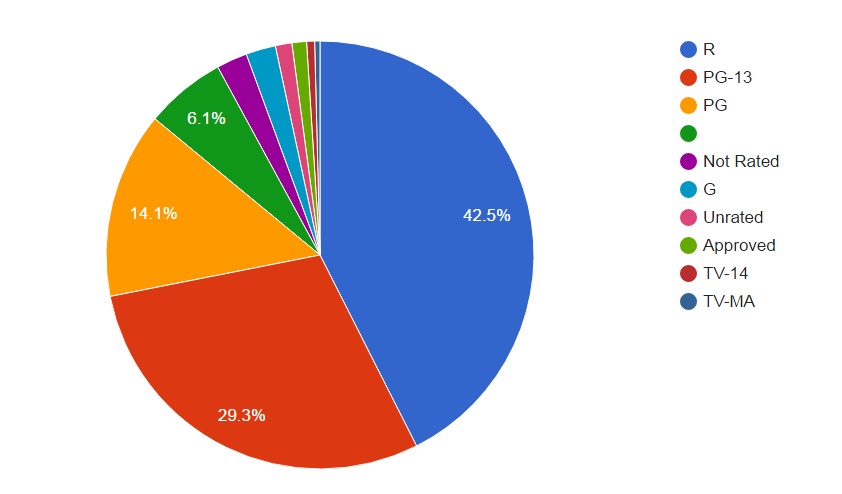
**Description of Attributes**

* Color – color of the movie (black and white or colored)
* Director – with more than 500 names
* Number of critic reviews
* Duration of the movies in minutes
* Director facebook likes
* Actor 1 , 2 , 3 names in different columns each
* Gross collected from the movie
* Genres – with a lot of different combinations
* Movie title
* Number of voted users on each movie in the page
* Total Facebook likes of the cast
* Facenumber in poster (faces included in main poster)
* Plot Keywords
* IMDB Movie links
* Languages
* Country of origin
* Content Rating
* Budget
* Title Year
* IMDB Score
* Aspect ratio
* Total of facebook likes which the movie has gathered

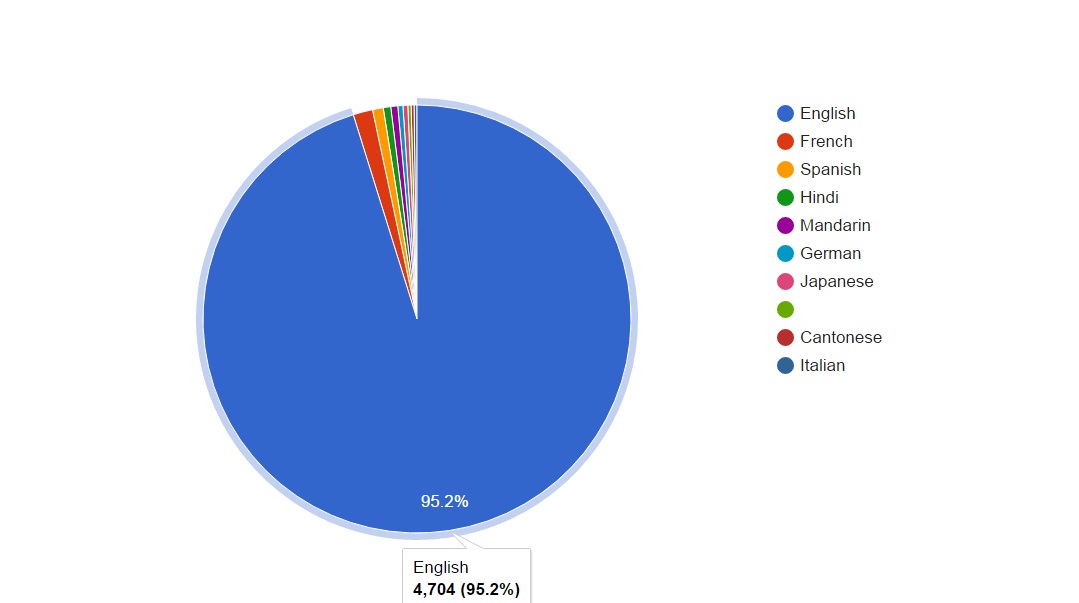
**Analysis and Visualisation**I decided to try and examine the most important part of the movies which is the amount of gross per budget gathered. As seen in the graph the early movies had a very small budget most ranging no more than a few millions for a big project and gross was not calculated very much back then. After year 1980 when the movies developed new technologies it was easier to gather more data. As we can see lately (after 2010) the movies have spiked with a huge amount of budget and only a few of them have returned in gross (no surprise there).

****  
The graph shows the number of score from 0 to 10 and the amount of critic reviews given on them. As we can see the number is way too much for the movies which have higher than average rating and especially the ones above 7.5

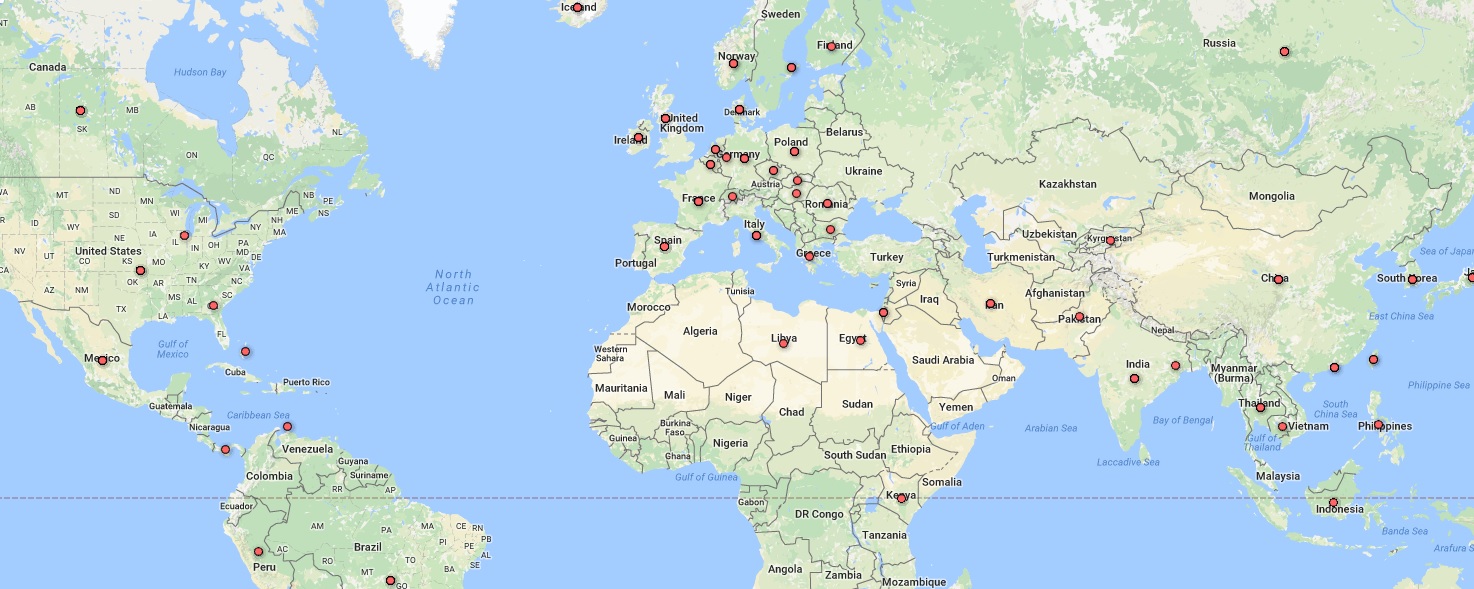
****

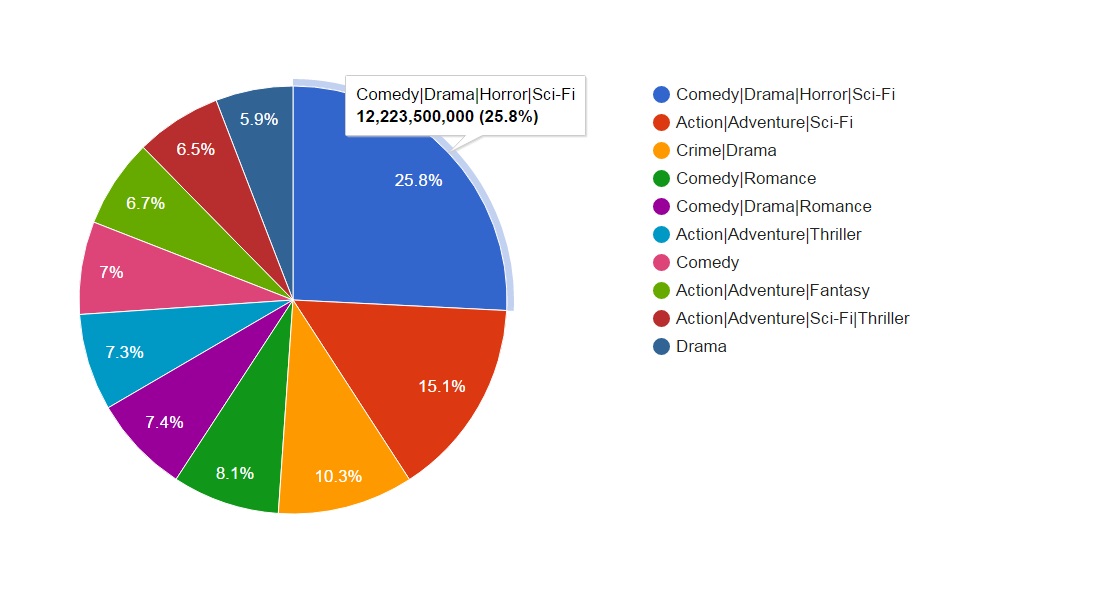
The rating of the gathered movies rated for different age of audience. The percent of “R” which stands for restricted type of movies has more than 40% which comes a little by surprise. Followed by more available to watch movies for kids from PG and PG13 with a total sum of 43%  
  


As most movies come from the United States with their huge industry and mostly in Hollywood with that percent raging to 95.2% as for sure that amount will be way less if all the international movies was included in the list.

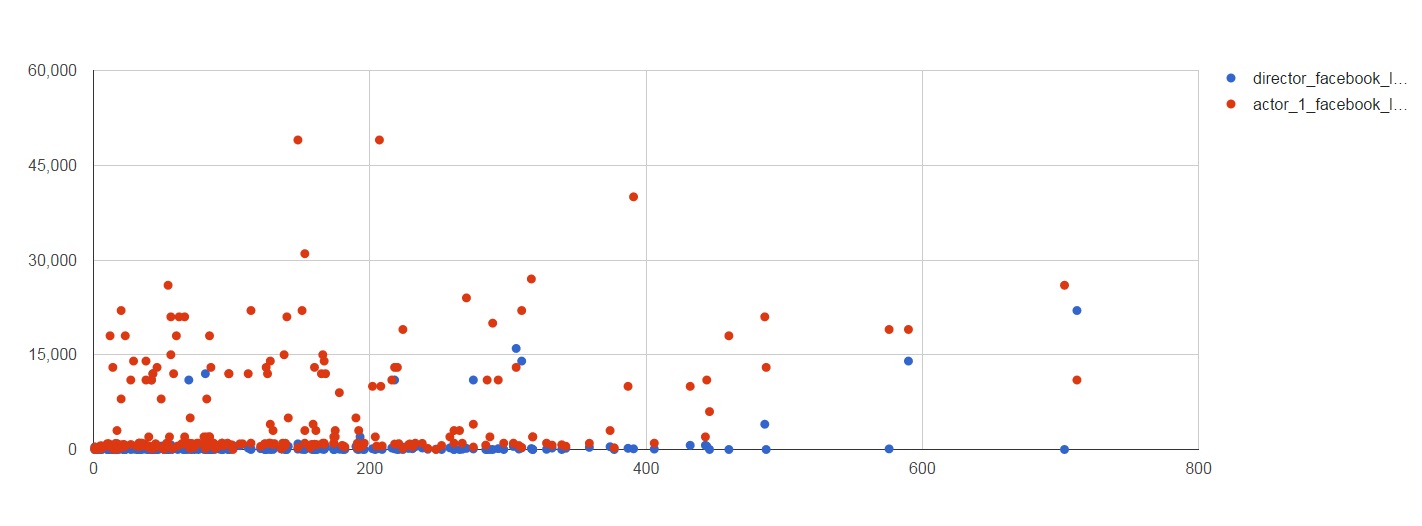
****

With the graph of languages was only appropriate to show the countries of origin and where the movies have been shot.

  
  
The data had a single problem, all the data had straight lines separating them rather than a comma which was really hard to visualize to either excel or google. But the data which was checked gave the result of the group of movies which had the tag of Comedy, Drama, Horror and Sci-Fi with an extraordinary and surprising result of almost 26%

  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
A graph showing the amount of likes for 1st 2nd and 3rd actor in each movie and its director. Showing that the 1st actor was favourited by the public and directors still being hardly recognized for his work.   


The director to 1st actor graph only

****

**Results and Findings**

In conclusion the given data shows that as the movie culture emerged it has been turning into profit and actors are getting more and more fame for their roles. Directors are still struggling to be remembered as much as them as they are the most important figure in making of each movie but they are staying behind the camera and don’t have any screen time. People like different and multicultural movies, having a lot of languages and mix of a few different kinds of genres rather than a simple comedy or just action packed movie.   
  
Most of the movies were recorded in the states but there is still more than 20% of those 5000+ which show that industry has moved to all continents and have recordings in all different countries ranging from language, economics and geological position. Most reviewed and commented movies were the ones with a rating of 7.5 and above as they got a lot more attention than the ones who didn’t leave any impression into the public and critics.

**Future Developments**

This data set can be used to see which movies had best results and maybe take a note on the cast, director and budget. Which is more or less like a mathematical formula or a recipe list. It’s only in the right amount of ingredients and equilibrium. It can be used to find patterns in which genre was good with what type of actors and how much revenue was gathered. Scores will show why the movies were so liked when compared with different results like budget and amount of Facebook fame and critics and fan reviews.   
The model can be expanded if the full amount of information was obtained and polished to work with better kind of software.

# References

(n.d.). Retrieved from https://www.truthinadvertising.org/id-theft-tops-complaints-2013/

*.amc.n*. (2016, 12 3).   
Retrieved from https://www.amc.nl/web/Zorg/Patient/Medisch-dossier/My-patient-record.htm

*Academy Datastax*. (2016, 12 1).  
 Retrieved from https://academy.datastax.com/planet-cassandra/nosql-performance-benchmarks

*anderson.ucla.edu*. (2016, 11 30). Retrieved from http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm

*Cassandra vs MongoDB*. (2016, 12 4). Retrieved from http://www.ippon.tech/blog/use-cassandra-mongodb-hbase-accumulo-mysql/

comments, A. f. (2016, 12 1). *Reddit*. Retrieved from https://www.reddit.com/r/AskReddit/comments/3o0cn6/serious\_victims\_of\_identity\_theft\_on\_reddit\_when/

*Google Fusion*. (2016, 12 06).   
Retrieved from https://fusiontables.google.com

*identitytheft*. (2016, 12 1).   
Retrieved from http://www.identitytheft.org.uk/

*KAGGLE*. (2016, 12 06).   
Retrieved from https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset

*Pensiero critico*. (2016, December 1).   
Retrieved from http://www.pensierocritico.eu/profilazione-identita-digitale.html

*The Modeling Agency*. (2016, 11 30).   
Retrieved from https://the-modeling-agency.com/how-data-mining-is-helping-healthcare/

University, C. C. (2016, 11 20). *Albodour, R. (2015) MongoDB Part 1 module 220CT*. Retrieved from Prezi: https://prezi.com/3ax4trxxq4z6/mongodb-part- 1/?utm\_campaign=share&utm\_medium=copy