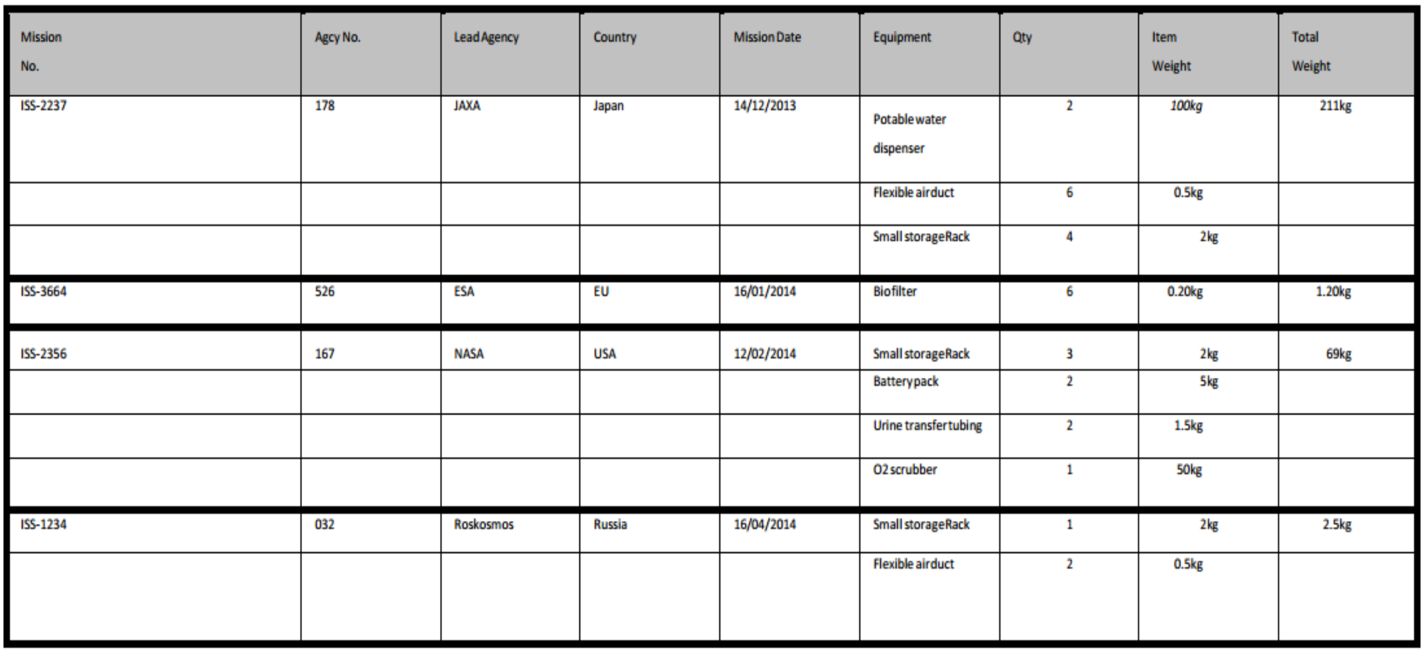
220CT – Data and Information Retrieval Coursework

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.



Mission No.	Agcy No.	Lead Agency	Country	Mission Date	Equipment	Qty	Item Weight	Total Weight
ISS-2237	178	JAXA	Japan	14/12/2013	Potable water dispenser	2	*100kg*	211kg
					Flexible airduct	6	0.5kg	
					Small storageRack	4	2kg	
ISS-3664	526	ESA	EU	16/01/2014	Biofilter	6	0.20kg	1.20kg
ISS-2356	167	NASA	USA	12/02/2014	Small storageRack	3	2kg	69kg
					Batterypack	2	5kg	
					Urine transfertubing	2	1.5kg	
					O2 scrubber	1	50kg	
ISS-1234	032	Roskosmos	Russia	16/04/2014	Small storageRack	1	2kg	2.5kg
					Flexible airduct	2	0.5kg	

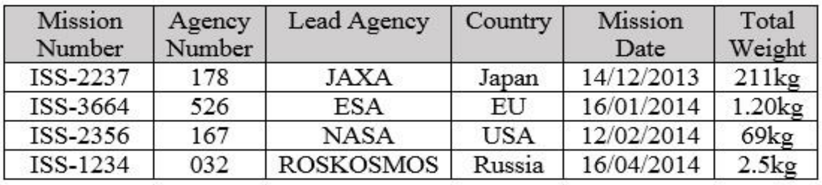
Deliverable: Using SQL, implement the database above. To do so, you'll need to normalise the table (should that be required), identify the attributes, create and Entity-Relationship Diagram and create the tables using SQL commands.

SOLUTION:

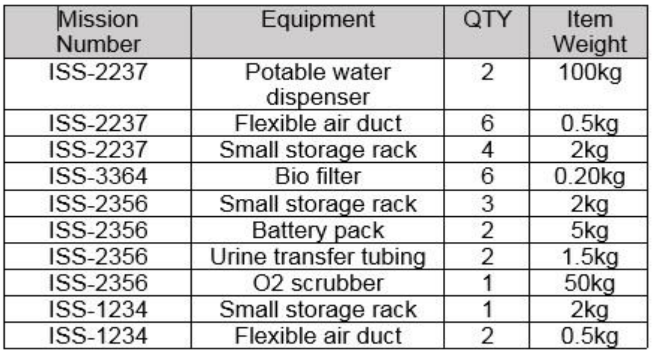
For this graph I would use traditional relation DB form because it's a small set of data entities and can easily be split to separate tables using normalisation.

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.



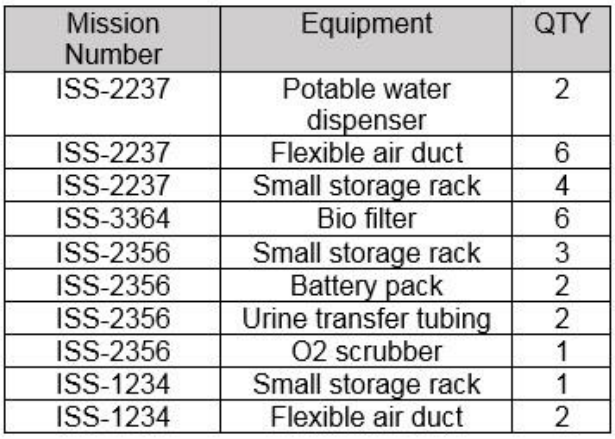
Mission Number	Agency Number	Lead Agency	Country	Mission Date	Total Weight
ISS-2237	178	JAXA	Japan	14/12/2013	211kg
ISS-3664	526	ESA	EU	16/01/2014	1.20kg
ISS-2356	167	NASA	USA	12/02/2014	69kg
ISS-1234	032	ROSKOSMOS	Russia	16/04/2014	2.5kg



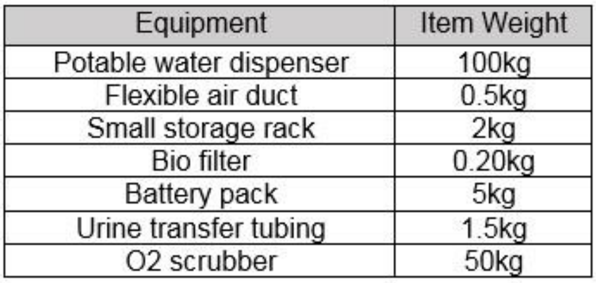
Mission Number	Equipment	QTY	Item Weight
ISS-2237	Potable water dispenser	2	100kg
ISS-2237	Flexible air duct	6	0.5kg
ISS-2237	Small storage rack	4	2kg
ISS-3364	Bio filter	6	0.20kg
ISS-2356	Small storage rack	3	2kg
ISS-2356	Battery pack	2	5kg
ISS-2356	Urine transfer tubing	2	1.5kg
ISS-2356	O2 scrubber	1	50kg
ISS-1234	Small storage rack	1	2kg
ISS-1234	Flexible air duct	2	0.5kg

2NF

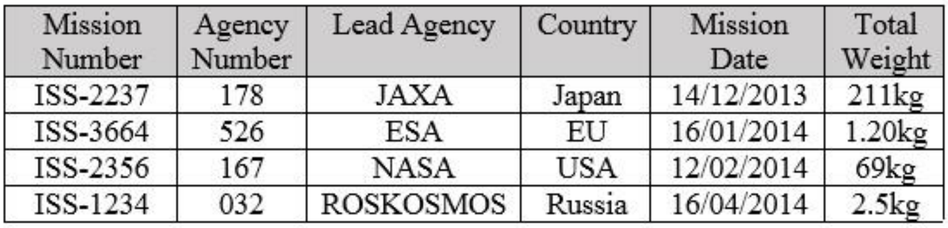
If I want it to be able to transform it from first form (1NF) into second form (2NF), any of the attributes depending only a part of a single table have to be removed and put into a third new one. Therefore the new table will hold the contents of equipment and weight from the “Equipment” table from 1NF. 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.



Mission Number	Equipment	QTY
ISS-2237	Potable water dispenser	2
ISS-2237	Flexible air duct	6
ISS-2237	Small storage rack	4
ISS-3364	Bio filter	6
ISS-2356	Small storage rack	3
ISS-2356	Battery pack	2
ISS-2356	Urine transfer tubing	2
ISS-2356	O2 scrubber	1
ISS-1234	Small storage rack	1
ISS-1234	Flexible air duct	2



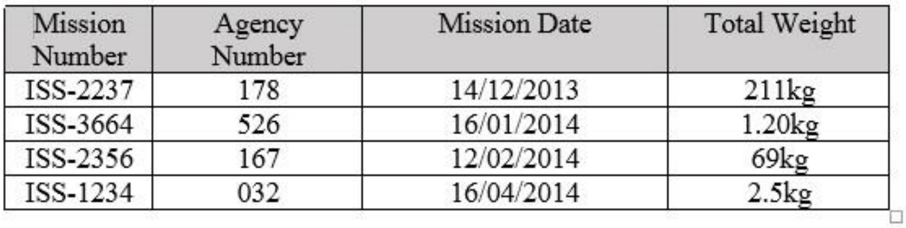
Equipment	Item Weight
Potable water dispenser	100kg
Flexible air duct	0.5kg
Small storage rack	2kg
Bio filter	0.20kg
Battery pack	5kg
Urine transfer tubing	1.5kg
O2 scrubber	50kg



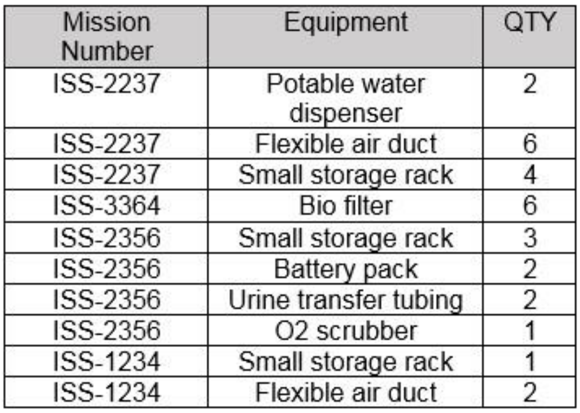
Mission Number	Agency Number	Lead Agency	Country	Mission Date	Total Weight
ISS-2237	178	JAXA	Japan	14/12/2013	211kg
ISS-3664	526	ESA	EU	16/01/2014	1.20kg
ISS-2356	167	NASA	USA	12/02/2014	69kg
ISS-1234	032	ROSKOSMOS	Russia	16/04/2014	2.5kg

3NF

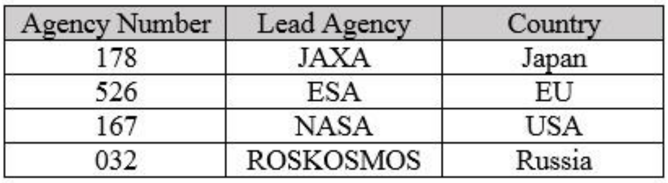
In order to transform from 2NF to 3NF, we have to remove any attributes that are more dependent on other non-key attributes and place them in a new table. The new table will be called “Agency”. Looking after that at the table “Equipment” we see that I has no primary key because there are no unique attributes,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



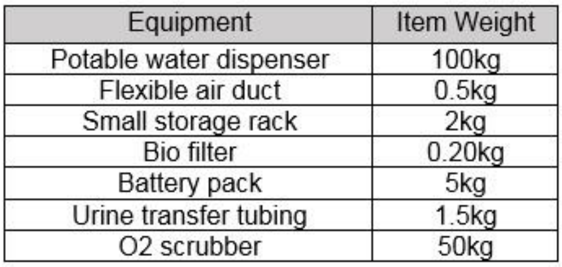
Mission Number	Agency Number	Mission Date	Total Weight
ISS-2237	178	14/12/2013	211kg
ISS-3664	526	16/01/2014	1.20kg
ISS-2356	167	12/02/2014	69kg
ISS-1234	032	16/04/2014	2.5kg



Mission Number	Equipment	QTY
ISS-2237	Potable water dispenser	2
ISS-2237	Flexible air duct	6
ISS-2237	Small storage rack	4
ISS-3364	Bio filter	6
ISS-2356	Small storage rack	3
ISS-2356	Battery pack	2
ISS-2356	Urine transfer tubing	2
ISS-2356	O2 scrubber	1
ISS-1234	Small storage rack	1
ISS-1234	Flexible air duct	2

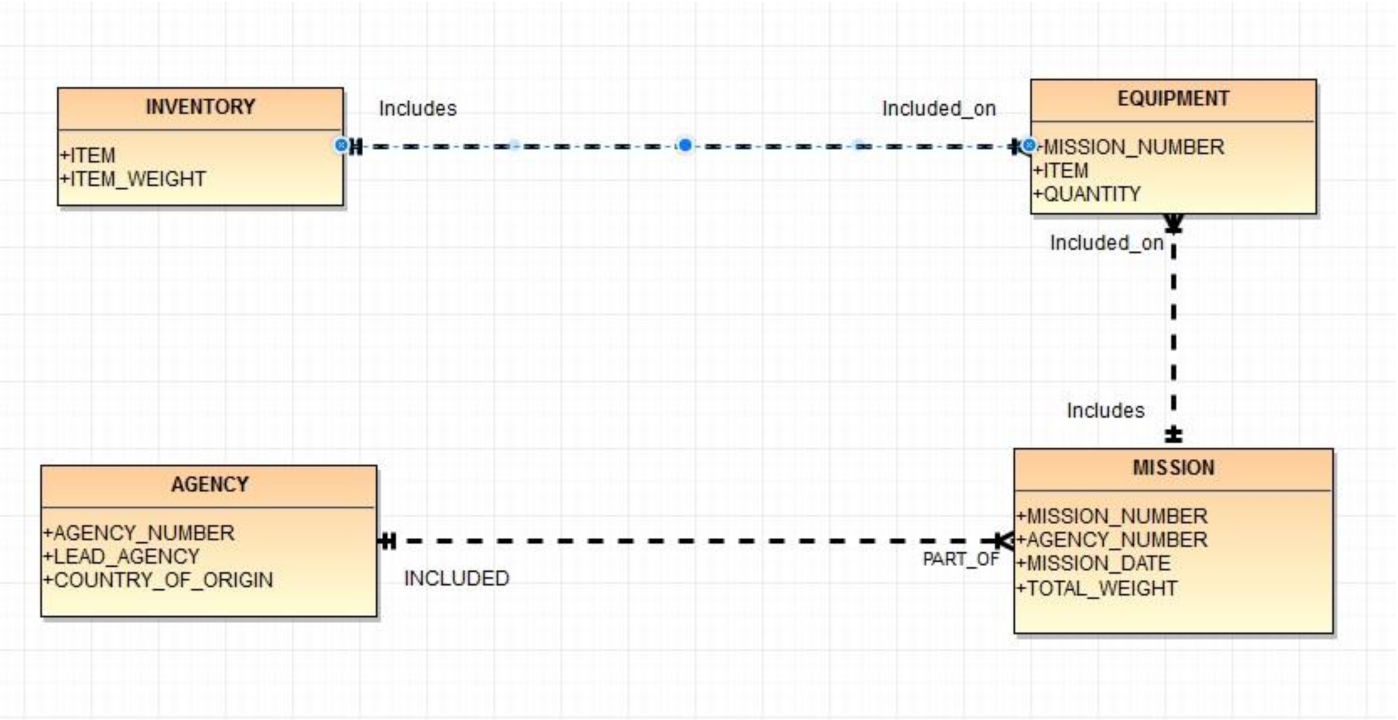


Agency Number	Lead Agency	Country
178	JAXA	Japan
526	ESA	EU
167	NASA	USA
032	ROSKOSMOS	Russia



Equipment	Item Weight
Potable water dispenser	100kg
Flexible air duct	0.5kg
Small storage rack	2kg
Bio filter	0.20kg
Battery pack	5kg
Urine transfer tubing	1.5kg
O2 scrubber	50kg

E-R DiagramThere are

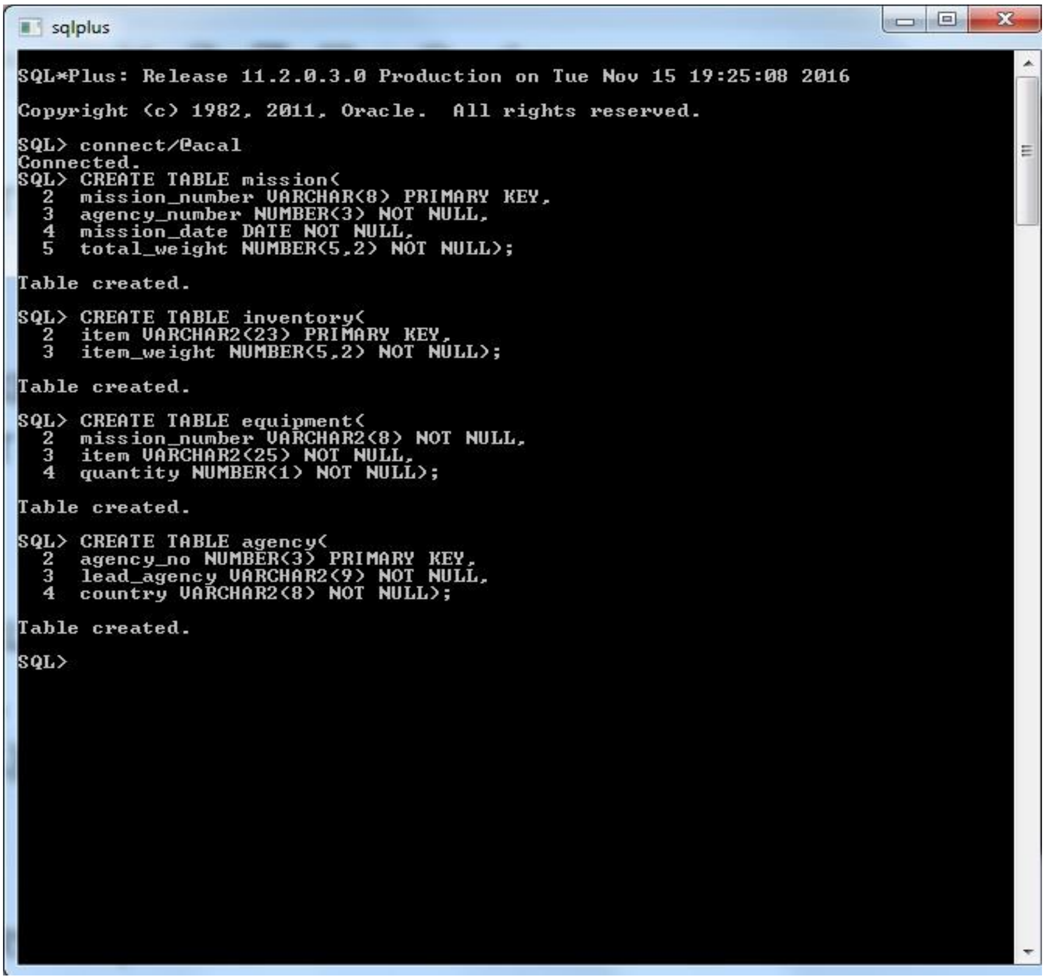


The four entities in this diagram. Agency, Inventory, Mission and Equipment share an entity relationship with each other as shown in the graph.

- 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”.

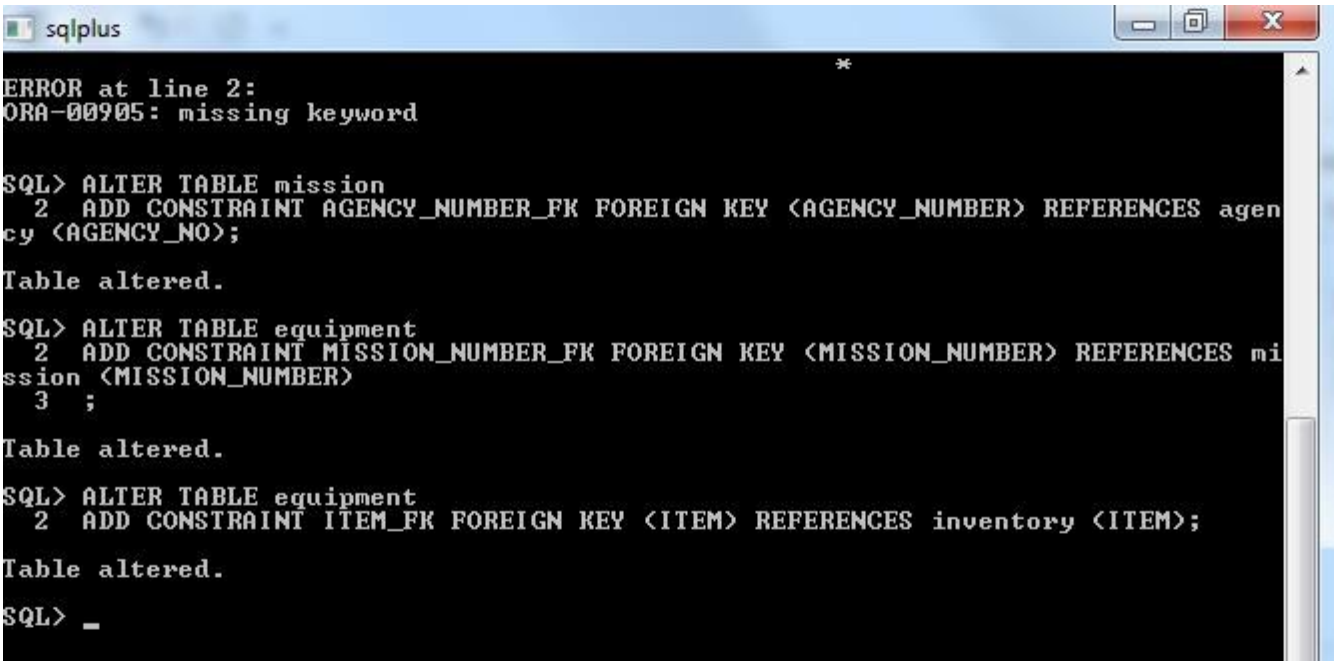
Implementing in SQL:

```
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);
```



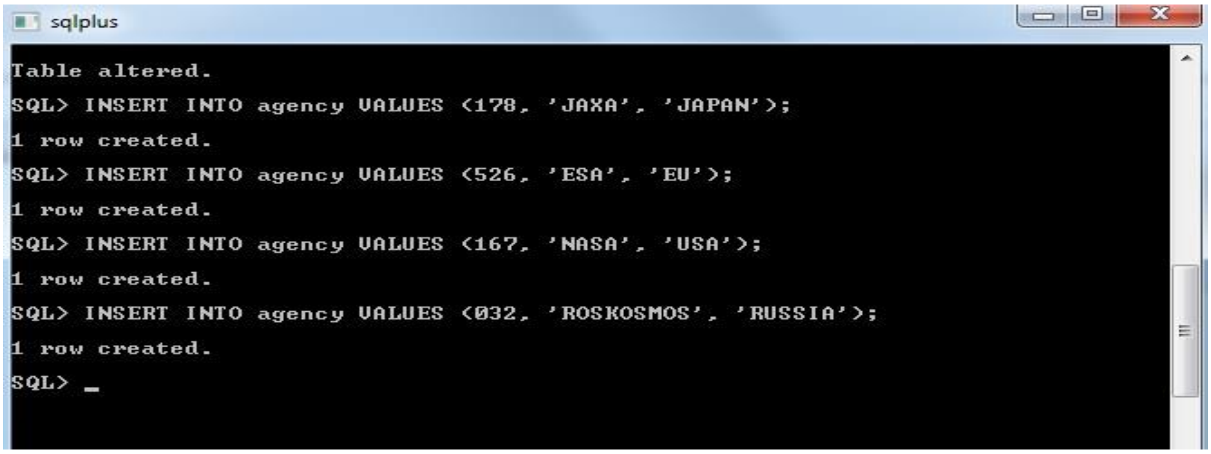
```
SQL*Plus: Release 11.2.0.3.0 Production on Tue Nov 15 19:25:08 2016

Copyright (c) 1982, 2011, Oracle.  All rights reserved.

SQL> connect/@acal
Connected.
SQL> CREATE TABLE mission(
  2  mission_number VARCHAR(8) PRIMARY KEY,
  3  agency_number NUMBER(3) NOT NULL,
  4  mission_date DATE NOT NULL,
  5  total_weight NUMBER(5,2) NOT NULL);

Table created.

SQL> CREATE TABLE inventory(
  2  item VARCHAR2(23) PRIMARY KEY,
  3  item_weight NUMBER(5,2) NOT NULL);

Table created.

SQL> CREATE TABLE equipment(
  2  mission_number VARCHAR2(8) NOT NULL,
  3  item VARCHAR2(25) NOT NULL,
  4  quantity NUMBER(1) NOT NULL);

Table created.

SQL> CREATE TABLE agency(
  2  agency_no NUMBER(3) PRIMARY KEY,
  3  lead_agency VARCHAR2(9) NOT NULL,
  4  country VARCHAR2(8) NOT NULL);

Table created.

SQL>
```

```
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);
```



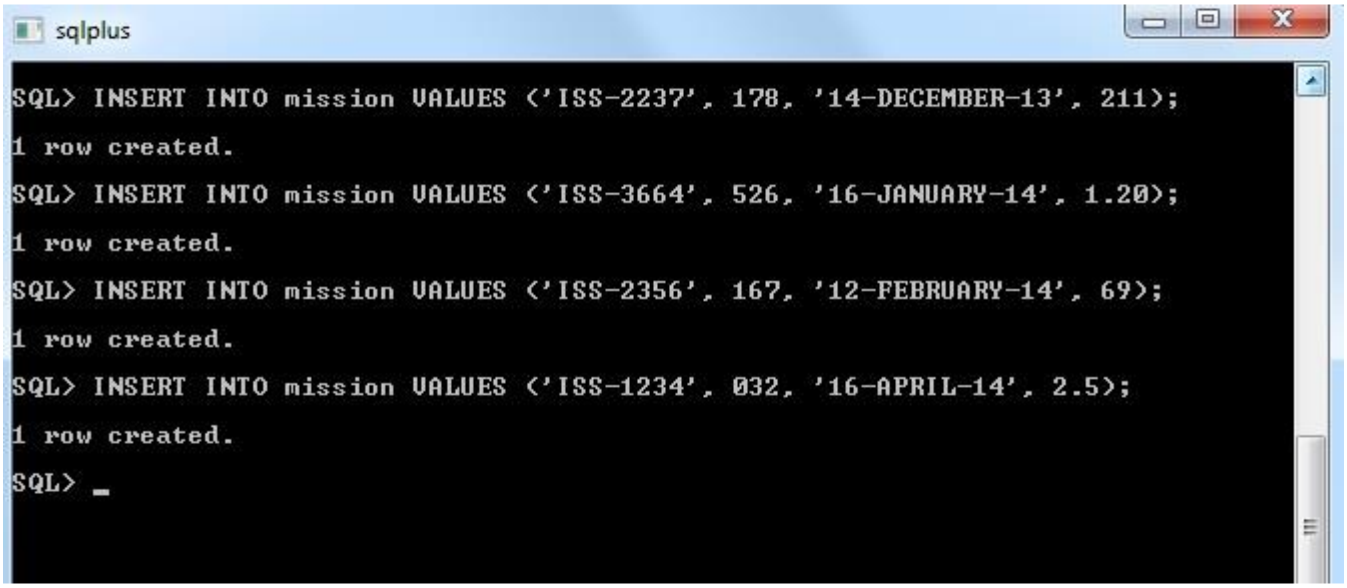
```
                                                            *
ERROR at line 2:
ORA-00905: missing keyword


SQL> ALTER TABLE mission
  2  ADD CONSTRAINT AGENCY_NUMBER_FK FOREIGN KEY (AGENCY_NUMBER) REFERENCES agen
cy (AGENCY_NO);

Table altered.

SQL> ALTER TABLE equipment
  2  ADD CONSTRAINT MISSION_NUMBER_FK FOREIGN KEY (MISSION_NUMBER) REFERENCES mi
ssion (MISSION_NUMBER)
  3  ;

Table altered.

SQL> ALTER TABLE equipment
  2  ADD CONSTRAINT ITEM_FK FOREIGN KEY (ITEM) REFERENCES inventory (ITEM);

Table altered.

SQL> _
```

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');



```
Table altered.
SQL> INSERT INTO agency VALUES (178, 'JAXA', 'JAPAN');
1 row created.
SQL> INSERT INTO agency VALUES (526, 'ESA', 'EU');
1 row created.
SQL> INSERT INTO agency VALUES (167, 'NASA', 'USA');
1 row created.
SQL> INSERT INTO agency VALUES (032, 'ROSKOSMOS', 'RUSSIA');
1 row created.
SQL> _
```

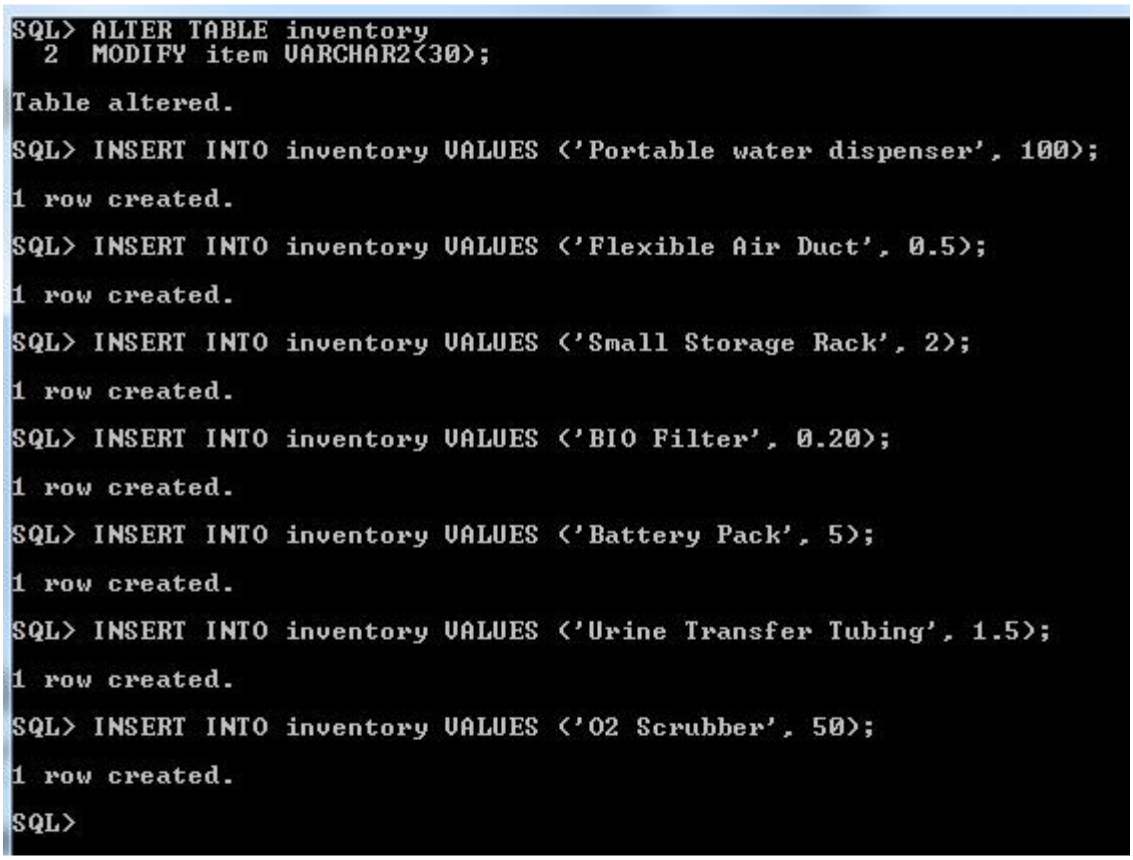
INSERT INTO mission VALUES ('ISS-2237', 178, '14-December-13', 211);
INSERT INTO mission VALUES ('ISS-3664', 526, '16-January-14', 1.20);
INSERT INTO mission VALUES ('ISS-2356', 167, '12-February-14', 69);
INSERT INTO mission VALUES ('ISS-1234', 032, '16-April-14', 2.5);



```
SQL> INSERT INTO mission VALUES (’ISS-2237’, 178, ’14-DECEMBER-13’, 211);

1 row created.

SQL> INSERT INTO mission VALUES (’ISS-3664’, 526, ’16-JANUARY-14’, 1.20);

1 row created.

SQL> INSERT INTO mission VALUES (’ISS-2356’, 167, ’12-FEBRUARY-14’, 69);

1 row created.

SQL> INSERT INTO mission VALUES (’ISS-1234’, 032, ’16-APRIL-14’, 2.5);

1 row created.

SQL> _
```

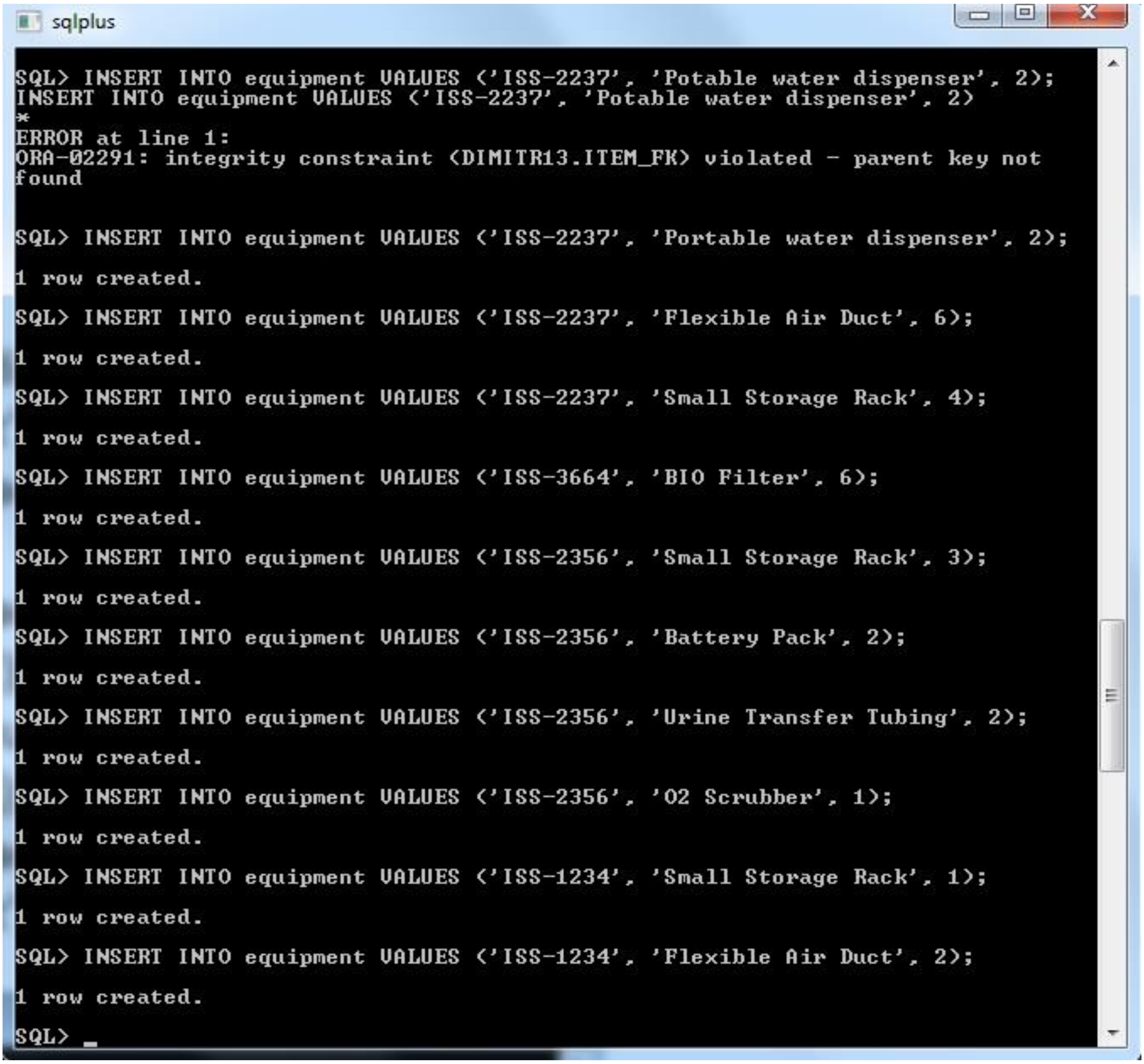
ALTER TABLE inventory
MODIFY item VARCHAR2 (30);
(*Saw that the first item didn’t have enough characters assigned so I modified it to accept a number of 30 characters)

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);



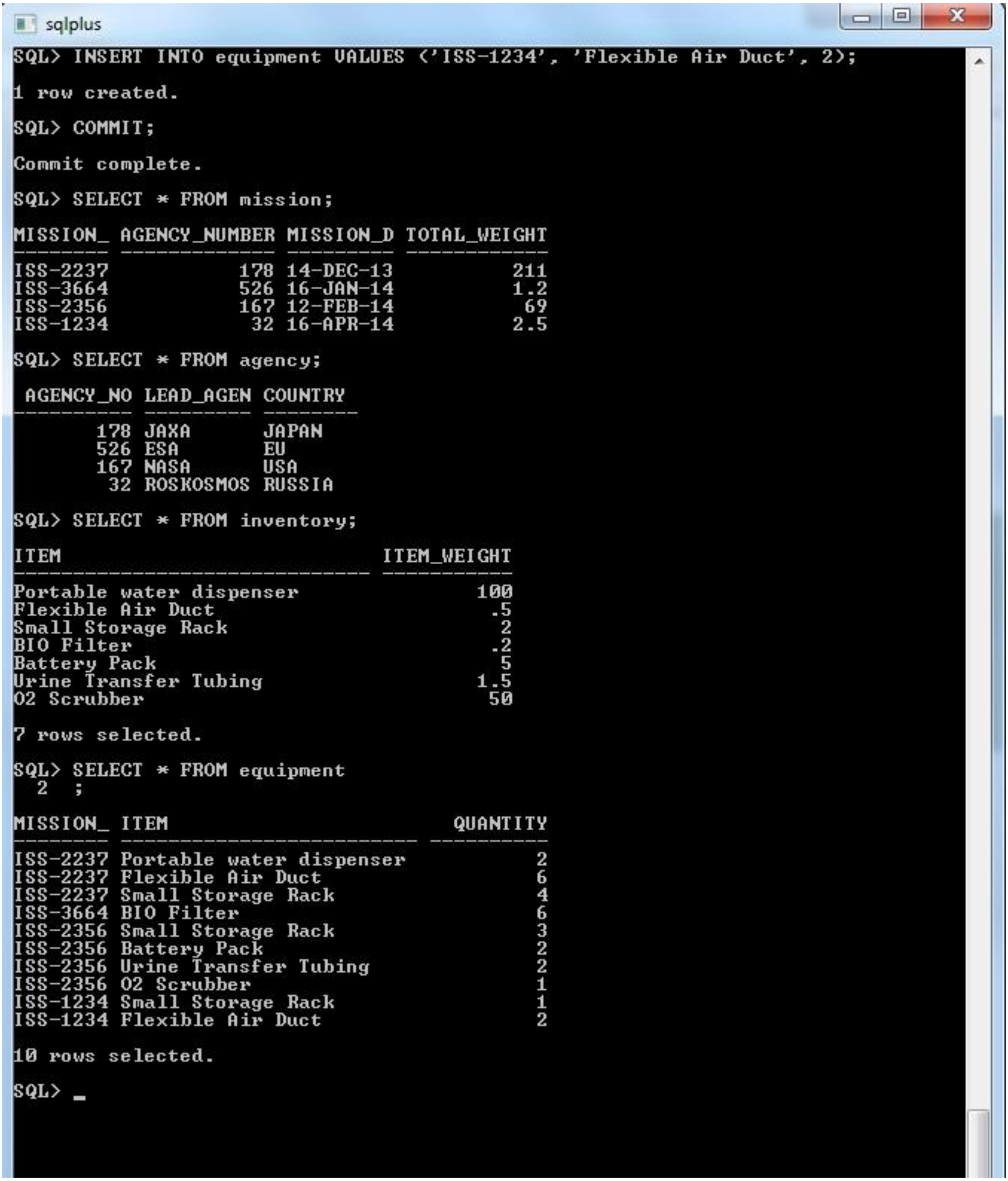
```
SQL> ALTER TABLE inventory
  2  MODIFY item VARCHAR2(30);

Table altered.

SQL> INSERT INTO inventory VALUES (’Portable water dispenser’, 100);

1 row created.

SQL> INSERT INTO inventory VALUES (’Flexible Air Duct’, 0.5);

1 row created.

SQL> INSERT INTO inventory VALUES (’Small Storage Rack’, 2);

1 row created.

SQL> INSERT INTO inventory VALUES (’BIO Filter’, 0.20);

1 row created.

SQL> INSERT INTO inventory VALUES (’Battery Pack’, 5);

1 row created.

SQL> INSERT INTO inventory VALUES (’Urine Transfer Tubing’, 1.5);

1 row created.

SQL> INSERT INTO inventory VALUES (’O2 Scrubber’, 50);

1 row created.

SQL>
```

```
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);
```



```
SQL> INSERT INTO equipment VALUES ('ISS-2237', 'Potable water dispenser', 2);
INSERT INTO equipment VALUES ('ISS-2237', 'Potable water dispenser', 2)
*
ERROR at line 1:
ORA-02291: integrity constraint (DIMITR13.ITEM_FK) violated - parent key not
found


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

1 row created.

SQL> INSERT INTO equipment VALUES ('ISS-2237', 'Flexible Air Duct', 6);

1 row created.

SQL> INSERT INTO equipment VALUES ('ISS-2237', 'Small Storage Rack', 4);

1 row created.

SQL> INSERT INTO equipment VALUES ('ISS-3664', 'BIO Filter', 6);

1 row created.

SQL> INSERT INTO equipment VALUES ('ISS-2356', 'Small Storage Rack', 3);

1 row created.

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

1 row created.

SQL> INSERT INTO equipment VALUES ('ISS-2356', 'Urine Transfer Tubing', 2);

1 row created.

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

1 row created.

SQL> INSERT INTO equipment VALUES ('ISS-1234', 'Small Storage Rack', 1);

1 row created.

SQL> INSERT INTO equipment VALUES ('ISS-1234', 'Flexible Air Duct', 2);

1 row created.

SQL> _
```

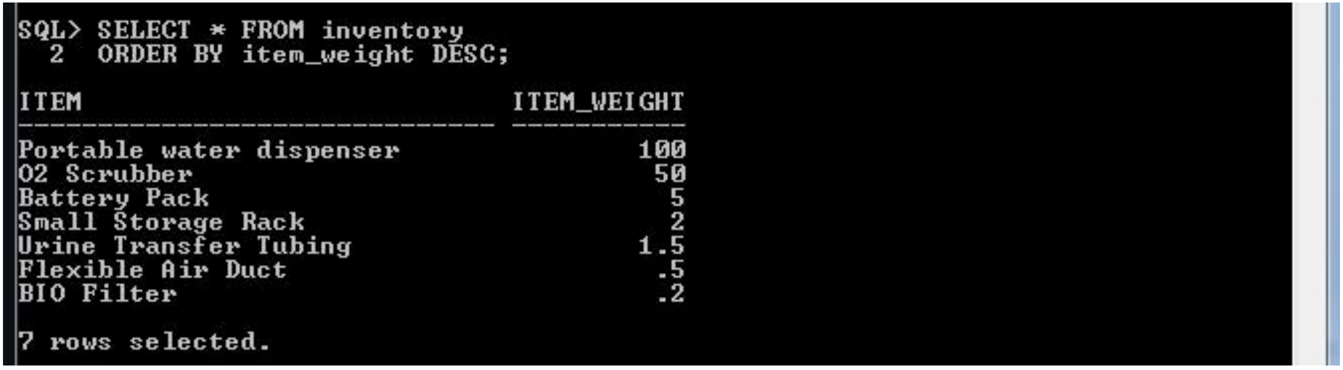
Commit and final look of the tables.



```
SQL> INSERT INTO equipment VALUES ('ISS-1234', 'Flexible Air Duct', 2);

1 row created.

SQL> COMMIT;

Commit complete.

SQL> SELECT * FROM mission;

MISSION_ AGENCY_NUMBER MISSION_D TOTAL_WEIGHT
-------- ------------- --------- ------------
ISS-2237           178 14-DEC-13          211
ISS-3664           526 16-JAN-14          1.2
ISS-2356           167 12-FEB-14           69
ISS-1234            32 16-APR-14          2.5

SQL> SELECT * FROM agency;

 AGENCY_NO LEAD_AGEN COUNTRY
---------- --------- --------
       178 JAXA      JAPAN
       526 ESA       EU
       167 NASA      USA
        32 ROSKOSMOS RUSSIA

SQL> SELECT * FROM inventory;

ITEM                           ITEM_WEIGHT
------------------------------ -----------
Portable water dispenser               100
Flexible Air Duct                       .5
Small Storage Rack                       2
BIO Filter                              .2
Battery Pack                             5
Urine Transfer Tubing                  1.5
O2 Scrubber                             50

7 rows selected.

SQL> SELECT * FROM equipment
  2  ;

MISSION_ ITEM                           QUANTITY
-------- ------------------------- ----------
ISS-2237 Portable water dispenser            2
ISS-2237 Flexible Air Duct                   6
ISS-2237 Small Storage Rack                  4
ISS-3664 BIO Filter                          6
ISS-2356 Small Storage Rack                  3
ISS-2356 Battery Pack                        2
ISS-2356 Urine Transfer Tubing               2
ISS-2356 O2 Scrubber                         1
ISS-1234 Small Storage Rack                  1
ISS-1234 Flexible Air Duct                   2

10 rows selected.

SQL> _
```

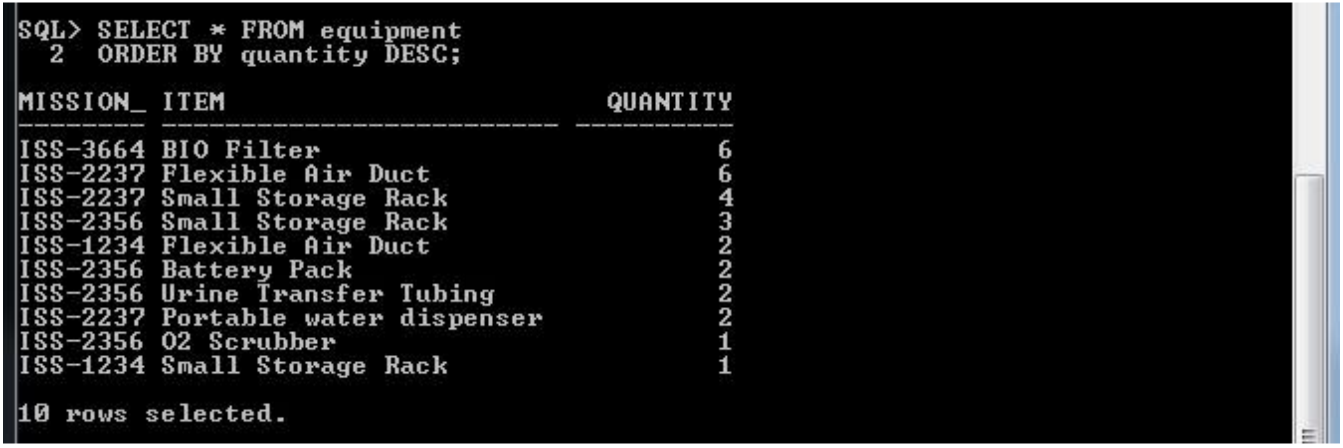
Examples for query with the working tables.

1. Produce an order of the items from the “inventory” table sorted in descending order of their weight.



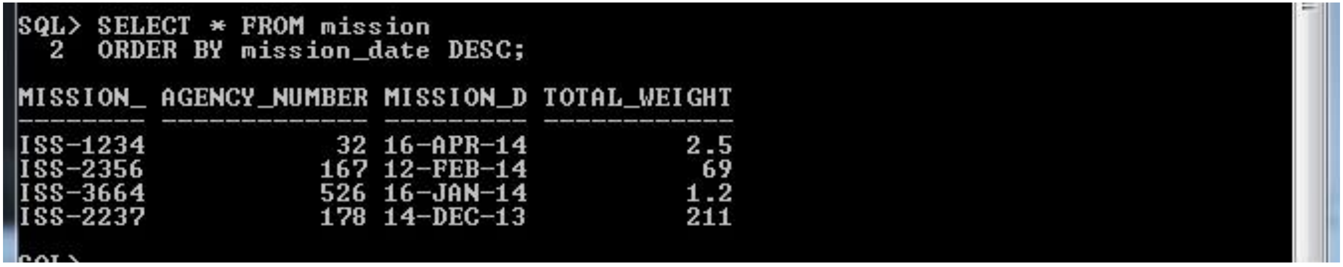
```
SQL> SELECT * FROM inventory
  2  ORDER BY item_weight DESC;

ITEM                           ITEM_WEIGHT
------------------------------ -----------
Portable water dispenser               100
O2 Scrubber                             50
Battery Pack                             5
Small Storage Rack                       2
Urine Transfer Tubing                  1.5
Flexible Air Duct                       .5
BIO Filter                              .2

7 rows selected.
```

2. Produce a list of the items in table “equipment” and sort them by their quantity for the acquired mission in descending order.



```
SQL> SELECT * FROM equipment
  2  ORDER BY quantity DESC;

MISSION_ ITEM                        QUANTITY
-------- ------------------------- ----------
ISS-3664 BIO Filter                         6
ISS-2237 Flexible Air Duct                  6
ISS-2237 Small Storage Rack                 4
ISS-2356 Small Storage Rack                 3
ISS-1234 Flexible Air Duct                  2
ISS-2356 Battery Pack                       2
ISS-2356 Urine Transfer Tubing              2
ISS-2237 Portable water dispenser           2
ISS-2356 O2 Scrubber                        1
ISS-1234 Small Storage Rack                 1

10 rows selected.
```

3. Produce a list of the items in table “mission”, ordered by their date of execution in descending order.



```
SQL> SELECT * FROM mission
  2  ORDER BY mission_date DESC;

MISSION_ AGENCY_NUMBER MISSION_D TOTAL_WEIGHT
-------- ------------- --------- ------------
ISS-1234            32 16-APR-14          2.5
ISS-2356           167 12-FEB-14           69
ISS-3664           526 16-JAN-14          1.2
ISS-2237           178 14-DEC-13          211

SQL>
```

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.

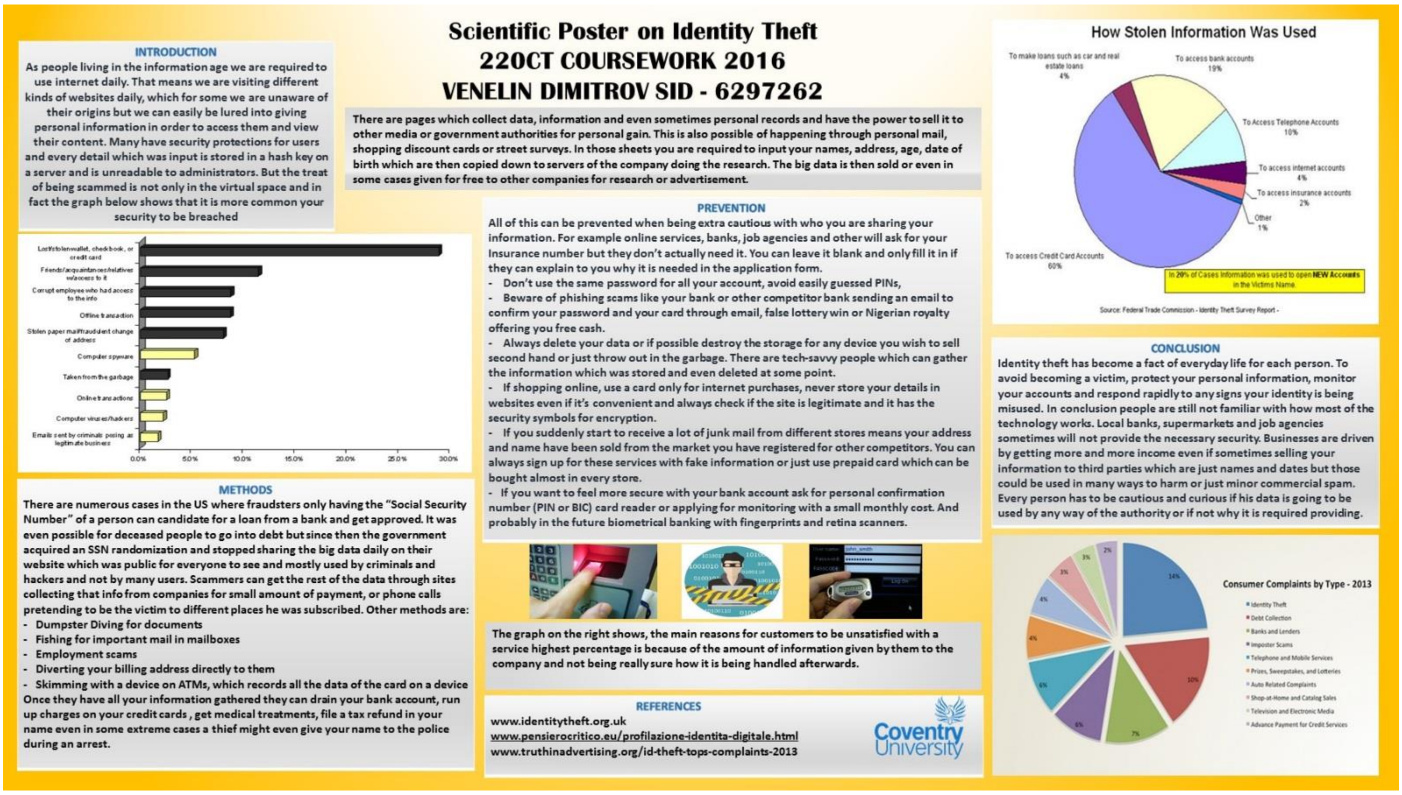
All of this can be prevented when being extra cautious with who you are sharing your information. For example online services, banks, job agencies and other will ask for your Insurance number but they don't actually need it. You can leave it blank and only fill it in if they can explain to you why it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



Scientific Poster on Identity Theft
220CT COURSEWORK 2016
VENELIN DIMITROV SID - 6297262

INTRODUCTION

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

Lost/stolen wallet, checkbook, or credit card
Friends/acquaintances/relatives w/access to it
Corrupt employee who had access to the info
Offline transaction
Stolen paper mail/fraudulent change of address
Computer spyware
Taken from the garbage
Online transactions
Computer viruses/hackers
Emails sent by criminals posing as legitimate business
0.0%
5.0%
10.0%
15.0%
20.0%
25.0%
30.0%

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.

METHODS

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 sites 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 ATMs, 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.

PREVENTION

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.

The graph on the right shows, the main reasons for customers to be unsatisfied with a service highest percentage is because of the amount of information given by them to the company and not being really sure how it is being handled afterwards.

REFERENCES

www.identitytheft.org.uk
www.pensierocritico.eu/profilazione-identita-digitale.html
www.truthinadvertising.org/id-theft-tops-complaints-2013

How Stolen Information Was Used

To make loans such as car and real estate loans 4%
To access bank accounts 19%
To Access Telephone Accounts 10%
To access internet accounts 4%
To access insurance accounts 2%
Other 1%
To access Credit Card Accounts 60%

In 20% of Cases Information was used to open **NEW Accounts** in the Victims Name

Source: Federal Trade Commission - Identity Theft Survey Report -

CONCLUSION

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.

Consumer Complaints by Type - 2013

14%
10%
7%
6%
6%
4%
4%
3%
3%
2%
Identity Theft
Debt Collection
Banks and Lenders
Imposter Scams
Telephone and Mobile Services
Prizes, Sweepstakes, and Lotteries
Auto Related Complaints
Shop-at-Home and Catalog Sales
Television and Electronic Media
Advance Payment for Credit Services

*The file is included in the submission form.

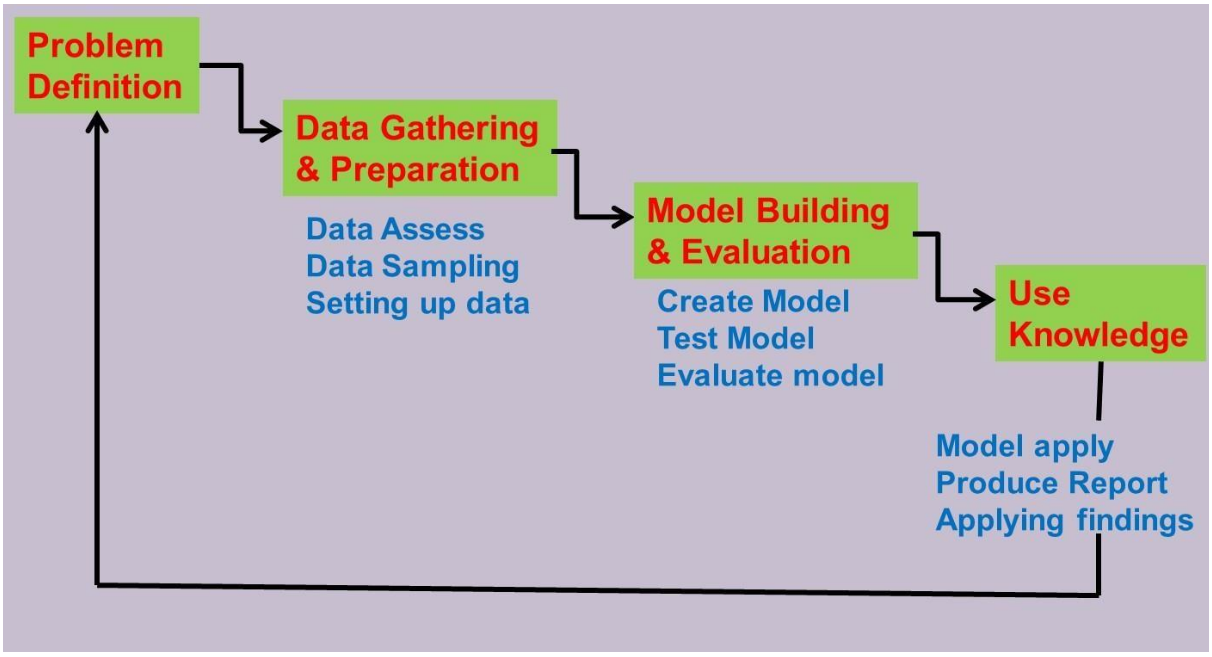
Task 3 – A data mining system for a Hospital (25 marks)

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.

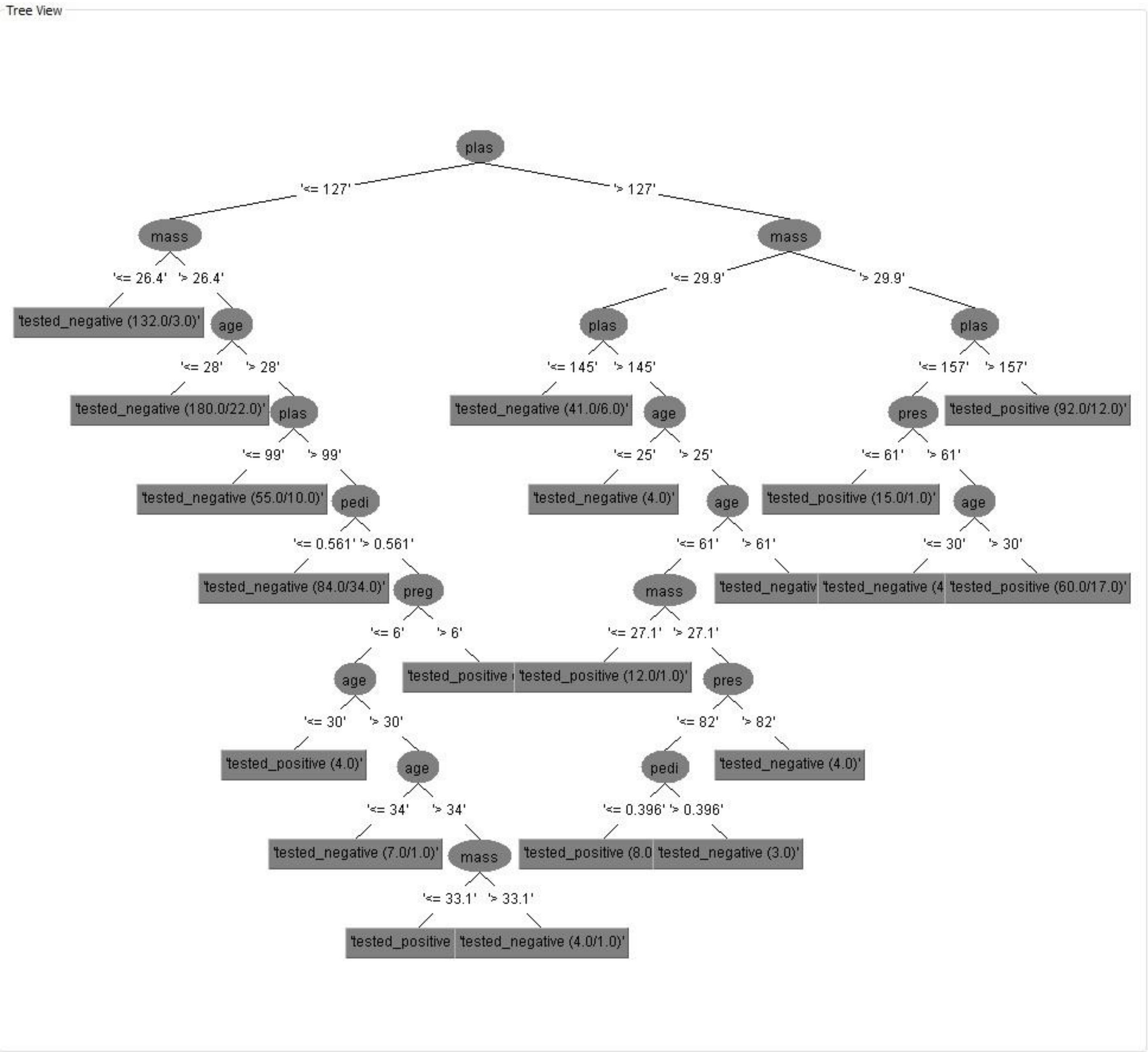
All the subjects in the list are females, with at least 21 years of age from PIMA Indian heritage

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

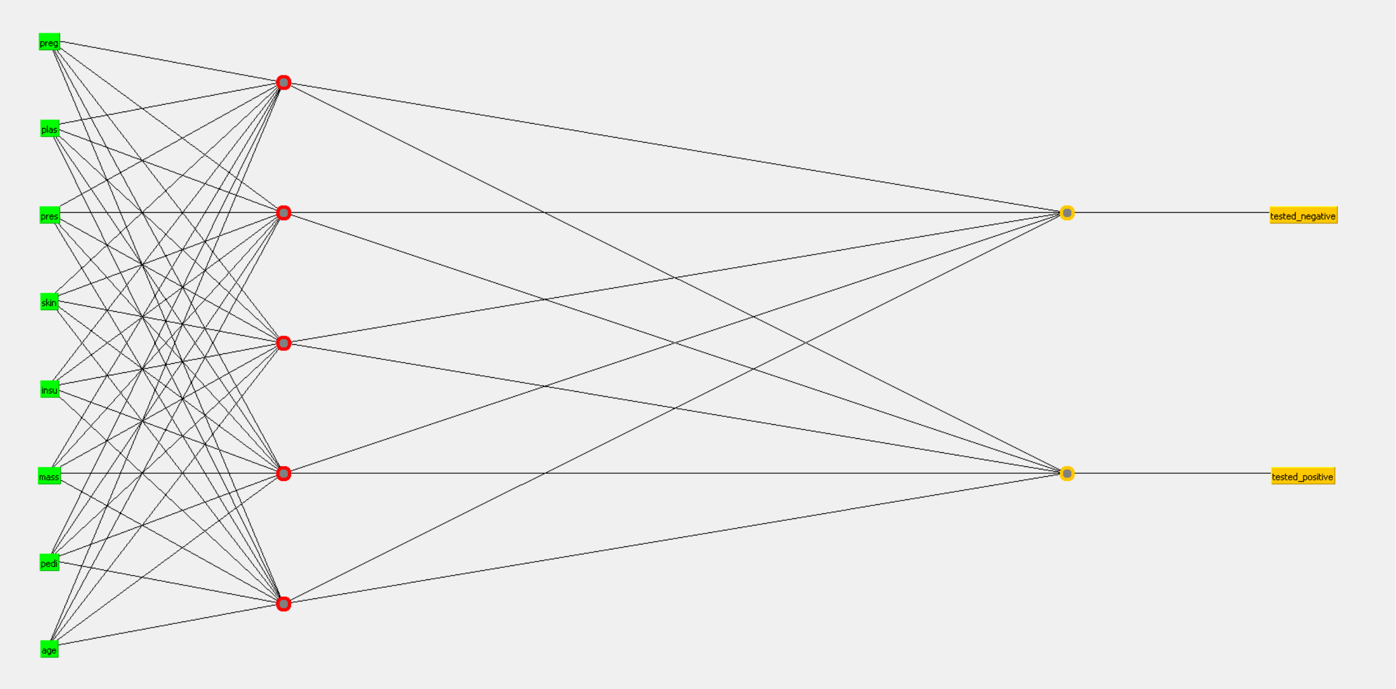


Summary

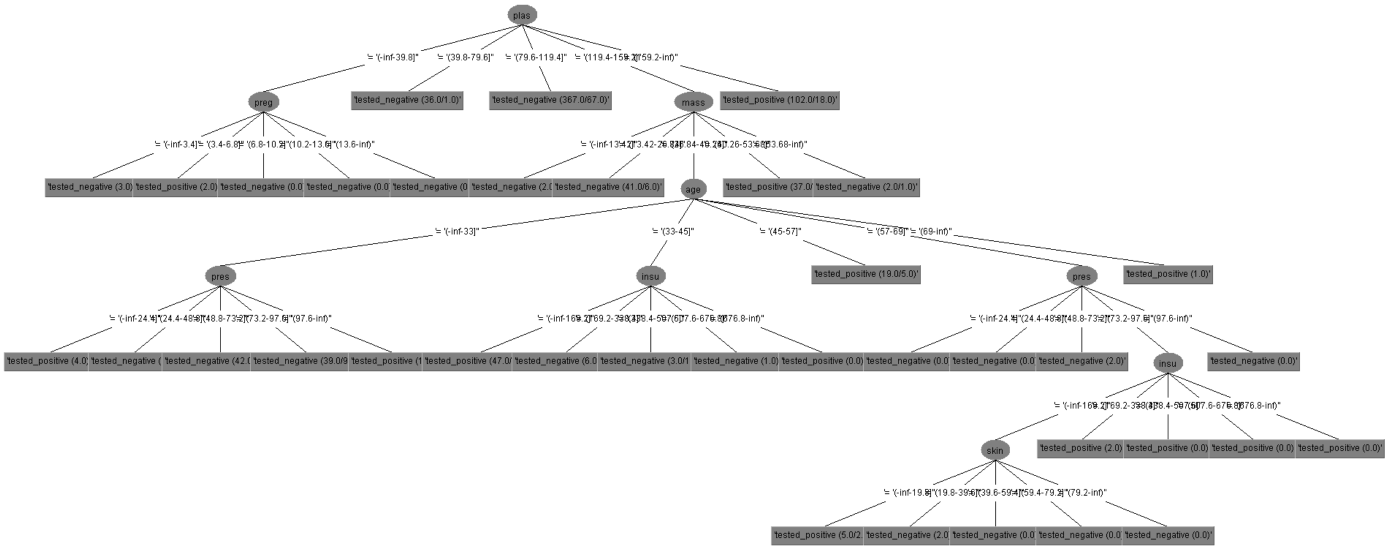
```
Correctly Classified Instances        567              73.8281 %
Incorrectly Classified Instances       201              26.1719 %
Number of total Instances = 768
```

```
. === 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.

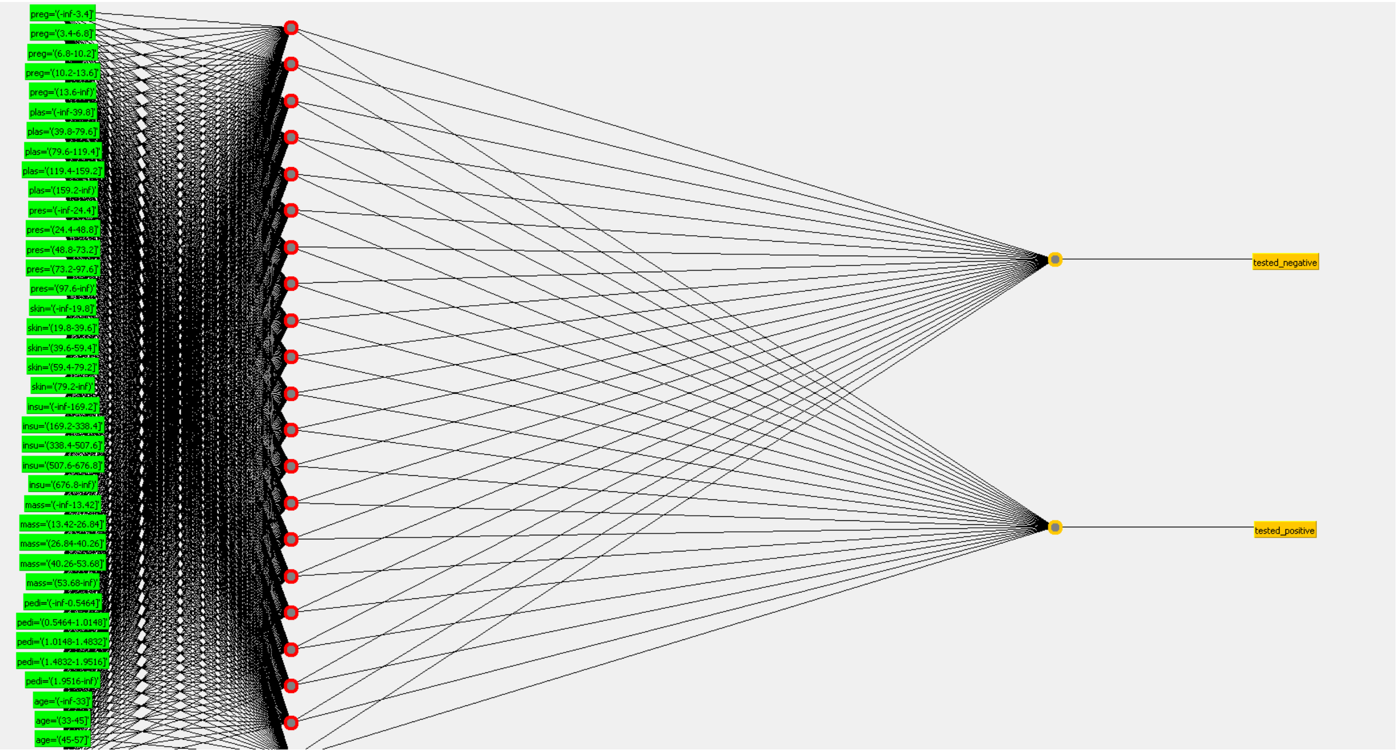


Decision Tree when Age is put into five different categories.



```
=== 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
```

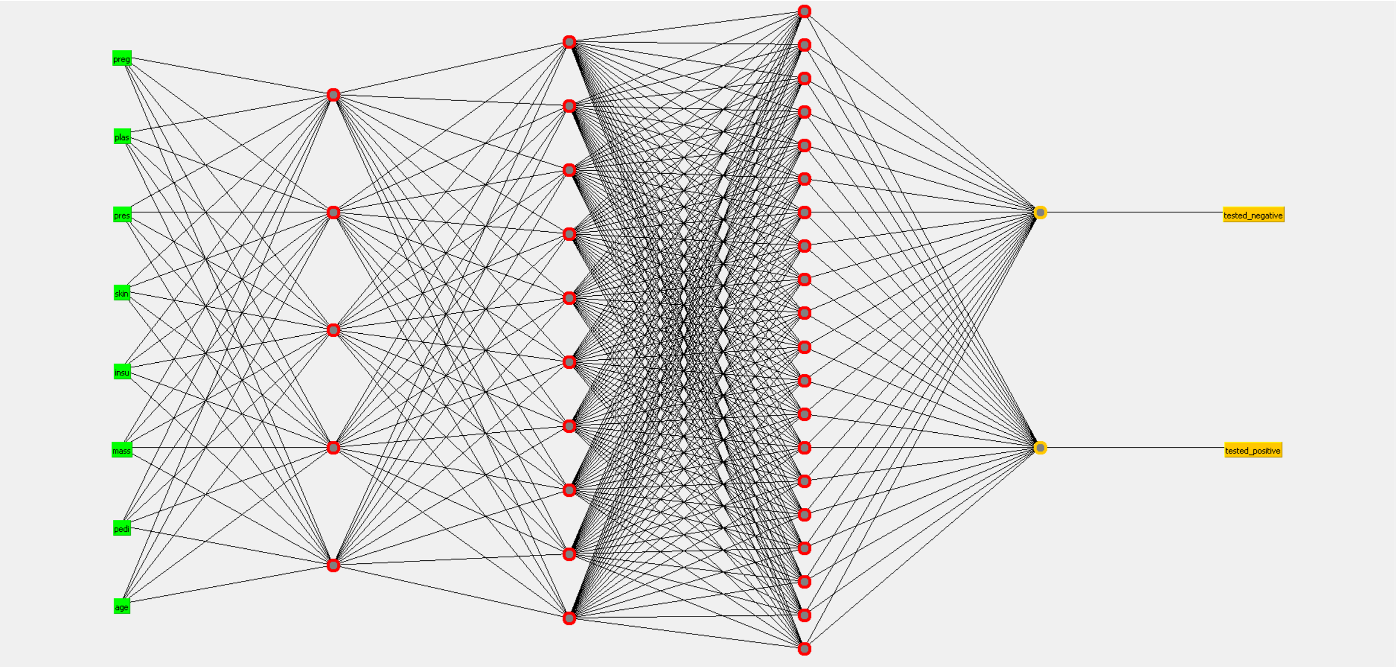
Based on this result 547 are tested “Negative” and 211 had the result as Positive.



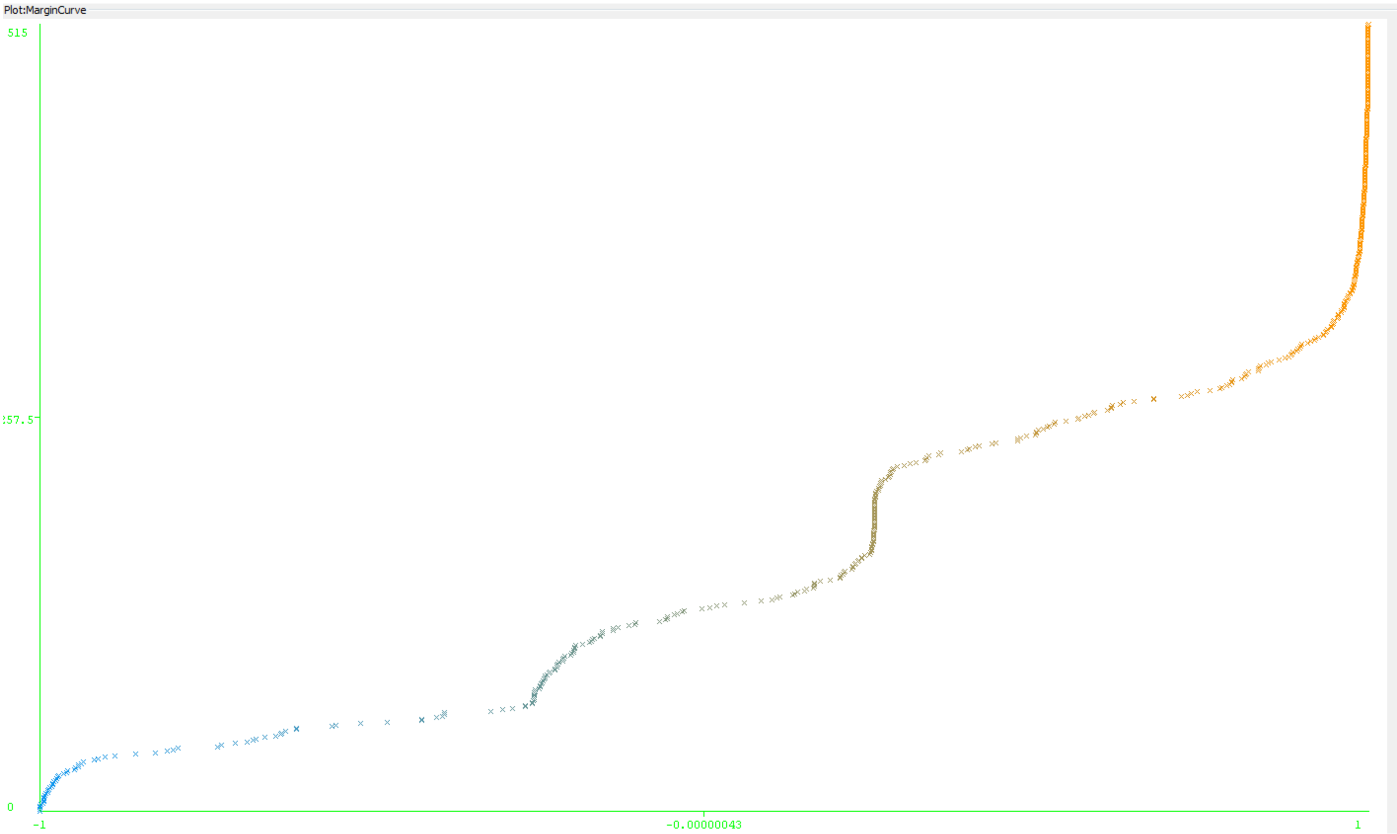
```
=== Summary ===
Correctly Classified Instances        541             73.8281 %
Incorrectly Classified Instances       227             26.1719 %
Total Number of Instances              768

=== Confusion Matrix ===
   a   b   <-- classified as
 371  95 |    a = tested_negative
 151 151 |    b = tested_positive
```

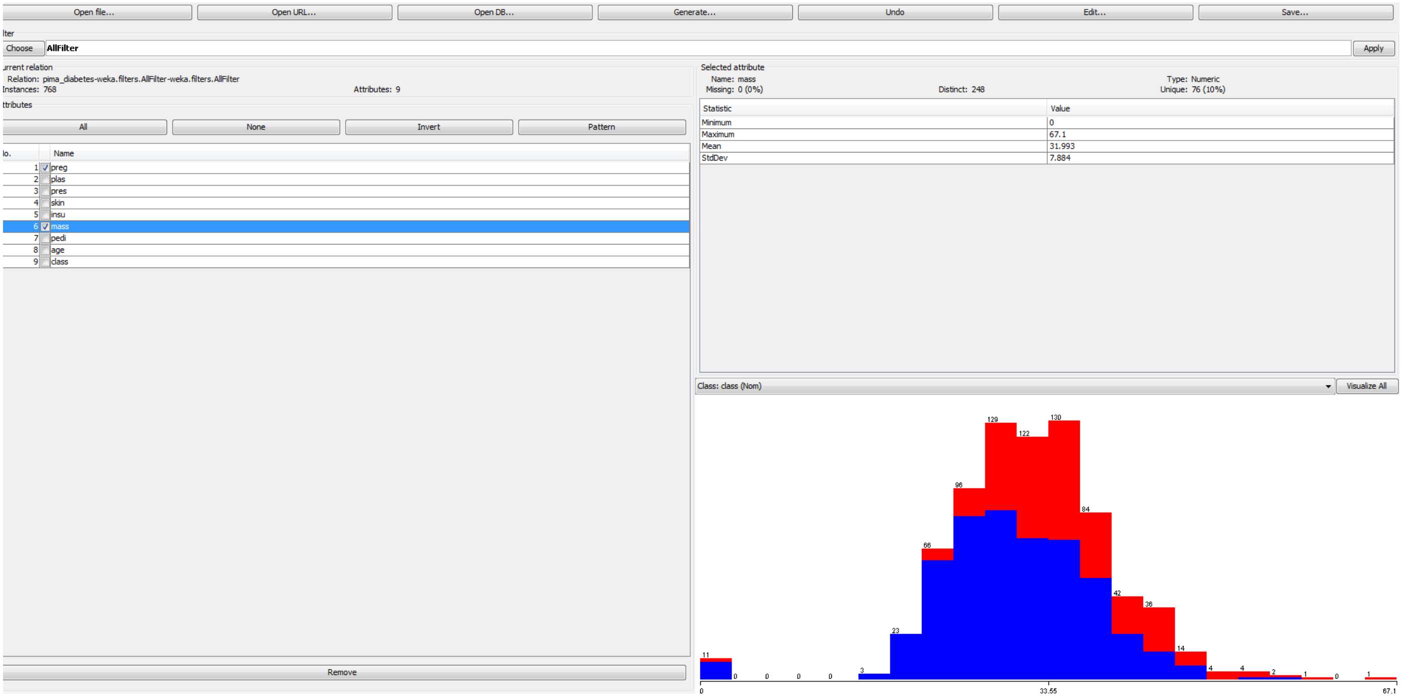
Based on this result 522 are tested “Negative” and 246 had the result as Positive.

<u>Hidden Layers Shown</u>



```
=== 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
```





lter

AllFilter

urrent relation

Relation: pima_diabetes-weka.filters.AllFilter-weka.filters.AllFilter

Instances: 768 Attributes: 9

ttributes

lo.		Name
1	☑	preg
2	□	plas
3	□	pres
4	□	skin
5	□	insu
6	☑	mass
7	□	pedi
8	□	age
9	□	class

Selected attribute

Name: mass Type: Numeric

Missing: 0 (0%) Distinct: 248 Unique: 76 (10%)

Statistic	Value
Minimum	0
Maximum	67.1
Mean	31.993
StdDev	7.884

Class: class (Nom)

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

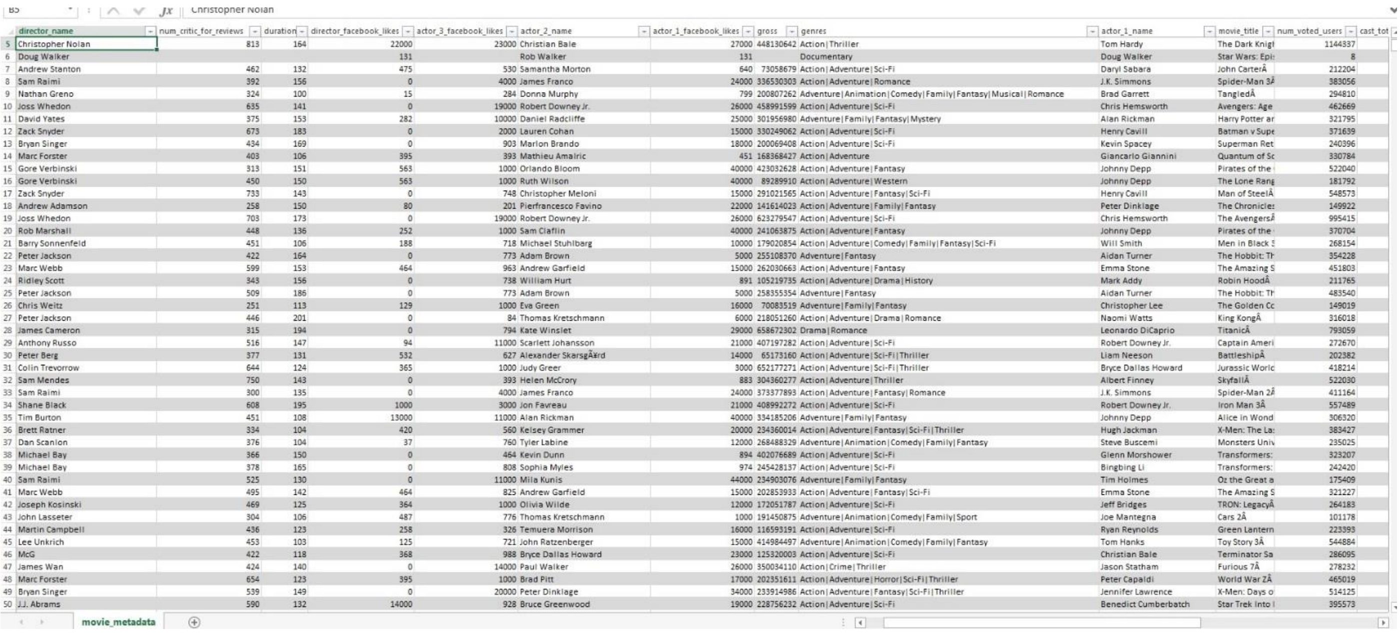
The creation of the movies started in the early 19th century when the motion picture cameras were invented and film production companies started to establish. One of the first movies date to year 1870s, back then the technology had only the power to record black and white pictures with around 100 shots 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.

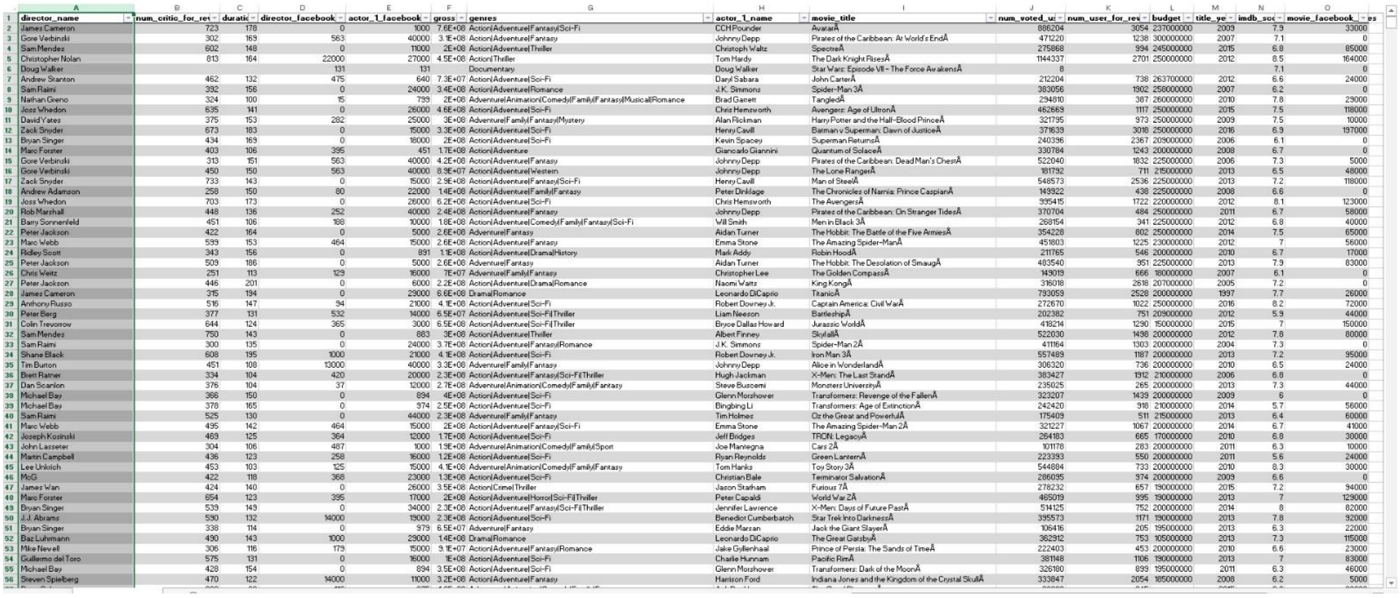
Acquiring the data

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.



	director_name	num_critic_for_reviews	duration	director_facebook_likes	actor_3_facebook_likes	actor_2_name	actor_1_facebook_likes	gross	genres	actor_1_name	movie_title	num_voted_users	cast_tot
5	Christopher Nolan	813	164	22000	23000	Christian Bale	27000	448130642	Action|Thriller	Tom Hardy	The Dark Knigl	1144337	
6	Doug Walker			131		Rob Walker	131		Documentary	Doug Walker	Star Wars: Epi:	8	
7	Andrew Stanton	462	132	475	530	Samantha Morton	640	73058679	Action|Adventure|Sci-Fi	Daryl Sabara	John CarterÂ	212204	
8	Sam Raimi	392	156	0	4000	James Franco	24000	336530303	Action|Adventure|Romance	J.K. Simmons	Spider-Man 3Â	383056	
9	Nathan Greno	324	100	15	284	Donna Murphy	799	200807262	Adventure|Animation|Comedy|Family|Fantasy|Musical|Romance	Brad Garrett	TangledÂ	294810	
10	Joss Whedon	635	141	0	19000	Robert Downey Jr.	26000	458991599	Action|Adventure|Sci-Fi	Chris Hemsworth	Avengers: Age	462669	
11	David Yates	375	153	282	10000	Daniel Radcliffe	25000	301956980	Adventure|Family|Fantasy|Mystery	Alan Rickman	Harry Potter ar	321795	
12	Zack Snyder	673	183	0	2000	Lauren Cohan	15000	330249062	Action|Adventure|Sci-Fi	Henry Cavill	Batman v Supe	371639	
13	Bryan Singer	434	169	0	903	Marlon Brando	18000	200069408	Action|Adventure|Sci-Fi	Kevin Spacey	Superman Ret	240396	
14	Marc Forster	403	106	395	393	Mathieu Amalric	451	168368427	Action|Adventure	Giancarlo Giannini	Quantum of Sc	330784	
15	Gore Verbinski	313	151	563	1000	Orlando Bloom	40000	423032628	Action|Adventure|Fantasy	Johnny Depp	Pirates of the	522040	
16	Gore Verbinski	450	150	563	1000	Ruth Wilson	40000	89289910	Action|Adventure|Western	Johnny Depp	The Lone Rang	181792	
17	Zack Snyder	733	143	0	748	Christopher Meloni	15000	291021565	Action|Adventure|Fantasy|Sci-Fi	Henry Cavill	Man of SteelÂ	548573	
18	Andrew Adamson	258	150	80	201	Pierfrancesco Favino	22000	141614023	Action|Adventure|Family|Fantasy	Peter Dinklage	The Chronicle:	149922	
19	Joss Whedon	703	173	0	19000	Robert Downey Jr.	26000	623279547	Action|Adventure|Sci-Fi	Chris Hemsworth	The AvengersÂ	995415	
20	Rob Marshall	448	136	252	1000	Sam Claflin	40000	241063875	Action|Adventure|Fantasy	Johnny Depp	Pirates of the	370704	
21	Barry Sonnenfeld	451	106	188	718	Michael Stuhlbarg	10000	179020854	Action|Adventure|Comedy|Family|Fantasy|Sci-Fi	Will Smith	Men in Black 3	268154	
22	Peter Jackson	422	164	0	773	Adam Brown	5000	255108370	Adventure|Fantasy	Aidan Turner	The Hobbit: Th	354228	
23	Marc Webb	599	153	464	963	Andrew Garfield	15000	262030663	Action|Adventure|Fantasy	Emma Stone	The Amazing S	451803	
24	Ridley Scott	343	156	0	738	William Hurt	891	105219735	Action|Adventure|Drama|History	Mark Addy	Robin HoodÂ	211765	
25	Peter Jackson	509	186	0	773	Adam Brown	5000	258355354	Adventure|Fantasy	Aidan Turner	The Hobbit: Th	483540	
26	Chris Weitz	251	113	129	1000	Eva Green	16000	70083519	Adventure|Family|Fantasy	Christopher Lee	The Golden Cc	149019	
27	Peter Jackson	446	201	0	84	Thomas Kretschmann	6000	218051260	Action|Adventure|Drama|Romance	Naomi Watts	King KongÂ	316018	
28	James Cameron	315	194	0	794	Kate Winslet	29000	658672302	Drama|Romance	Leonardo DiCaprio	TitanicÂ	793059	
29	Anthony Russo	516	147	94	11000	Scarlett Johansson	21000	407197282	Action|Adventure|Sci-Fi	Robert Downey Jr.	Captain Ameri	272670	
30	Peter Berg	377	131	532	627	Alexander SkarsgÃ¥rd	14000	65173160	Action|Adventure|Sci-Fi|Thriller	Liam Neeson	BattleshipÂ	202382	
31	Colin Trevorrow	644	124	365	1000	Judy Greer	3000	652177271	Action|Adventure|Sci-Fi|Thriller	Bryce Dallas Howard	Jurassic Worlc	418214	
32	Sam Mendes	750	143	0	393	Helen McCrory	883	304360277	Action|Adventure|Thriller	Albert Finney	SkyfallÂ	522030	
33	Sam Raimi	300	135	0	4000	James Franco	24000	373377893	Action|Adventure|Fantasy|Romance	J.K. Simmons	Spider-Man 2Â	411164	
34	Shane Black	608	195	1000	3000	Jon Favreau	21000	408992272	Action|Adventure|Sci-Fi	Robert Downey Jr.	Iron Man 3Â	557489	
35	Tim Burton	451	108	13000	11000	Alan Rickman	40000	334185206	Adventure|Family|Fantasy	Johnny Depp	Alice in Wond	306320	
36	Brett Ratner	334	104	420	560	Kelsey Grammer	20000	234360014	Action|Adventure|Fantasy|Sci-Fi|Thriller	Hugh Jackman	X-Men: The La:	383427	
37	Dan Scanlon	376	104	37	760	Tyler Labine	12000	268488329	Adventure|Animation|Comedy|Family|Fantasy	Steve Buscemi	Monsters Univ	235025	
38	Michael Bay	366	150	0	464	Kevin Dunn	894	402076689	Action|Adventure|Sci-Fi	Glenn Morshower	Transformers:	323207	
39	Michael Bay	378	165	0	808	Sophia Myles	974	245428137	Action|Adventure|Sci-Fi	Bingbing Li	Transformers:	242420	
40	Sam Raimi	525	130	0	11000	Mila Kunis	44000	234903076	Adventure|Family|Fantasy	Tim Holmes	Oz the Great a	175409	
41	Marc Webb	495	142	464	825	Andrew Garfield	15000	202853933	Action|Adventure|Fantasy|Sci-Fi	Emma Stone	The Amazing S	321227	
42	Joseph Kosinski	469	125	364	1000	Olivia Wilde	12000	172051787	Action|Adventure|Sci-Fi	Jeff Bridges	TRON: LegacyÂ	264183	
43	John Lasseter	304	106	487	776	Thomas Kretschmann	1000	191450875	Adventure|Animation|Comedy|Family|Sport	Joe Mantegna	Cars 2Â	101178	
44	Martin Campbell	436	123	258	326	Temuera Morrison	16000	116593191	Action|Adventure|Sci-Fi	Ryan Reynolds	Green Lantern	223393	
45	Lee Unkrich	453	103	125	721	John Ratzenberger	15000	414984497	Adventure|Animation|Comedy|Family|Fantasy	Tom Hanks	Toy Story 3Â	544884	
46	McG	422	118	368	988	Bryce Dallas Howard	23000	125320003	Action|Adventure|Sci-Fi	Christian Bale	Terminator Sa	286095	
47	James Wan	424	140	0	14000	Paul Walker	26000	350034110	Action|Crime|Thriller	Jason Statham	Furious 7Â	278232	
48	Marc Forster	654	123	395	1000	Brad Pitt	17000	202351611	Action|Adventure|Horror|Sci-Fi|Thriller	Peter Capaldi	World War ZÂ	465019	
49	Bryan Singer	539	149	0	20000	Peter Dinklage	34000	233914986	Action|Adventure|Fantasy|Sci-Fi|Thriller	Jennifer Lawrence	X-Men: Days o	514125	
50	J.J. Abrams	590	132	14000	928	Bruce Greenwood	19000	228756232	Action|Adventure|Sci-Fi	Benedict Cumberbatch	Star Trek Into I	395573	

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.



1	director_name	num_critic_for_re	duratio	director_facebook	actor_1_facebook	gross	genres	actor_1_name	movie_title	num_voted_u	num_user_for_rev	budget	title_ye	imdb_sco	movie_facebook_
2	James Cameron	723	178	0	1000	7.6E+08	Action|Adventure|Fantasy|Sci-Fi	CCH Pounder	AvatarÂ	886204	3054	237000000	2009	7.9	33000
3	Gore Verbinski	302	169	563	40000	3.1E+08	Action|Adventure|Fantasy	Johnny Depp	Pirates of the Caribbean: At World's EndÂ	471220	1238	300000000	2007	7.1	0
4	Sam Mendes	602	148	0	11000	2E+08	Action|Adventure|Thriller	Christoph Waltz	SpectreÂ	275868	994	245000000	2015	6.8	85000
5	Christopher Nolan	813	164	22000	27000	4.5E+08	Action|Thriller	Tom Hardy	The Dark Knight RisesÂ	1144337	2701	250000000	2012	8.5	164000
6	Doug Walker			131	131		Documentary	Doug Walker	Star Wars: Episode VII - The Force AwakensÂ	8				7.1	0
7	Andrew Stanton	462	132	475	640	7.3E+07	Action|Adventure|Sci-Fi	Daryl Sabara	John CarterÂ	212204	738	263700000	2012	6.6	24000
8	Sam Raimi	392	156	0	24000	3.4E+08	Action|Adventure|Romance	J.K. Simmons	Spider-Man 3Â	383056	1902	258000000	2007	6.2	0
9	Nathan Greno	324	100	15	799	2E+08	Adventure|Animation|Comedy|Family|Fantasy|Musical|Romance	Brad Garrett	TangledÂ	294810	387	260000000	2010	7.8	29000
10	Joss Whedon	635	141	0	26000	4.6E+08	Action|Adventure|Sci-Fi	Chris Hemsworth	Avengers: Age of UltronÂ	462669	1117	250000000	2015	7.5	118000
11	David Yates	375	153	282	25000	3E+08	Adventure|Family|Fantasy|Mystery	Alan Rickman	Harry Potter and the Half-Blood PrinceÂ	321795	973	250000000	2009	7.5	10000
12	Zack Snyder	673	183	0	15000	3.3E+08	Action|Adventure|Sci-Fi	Henry Cavill	Batman v Superman: Dawn of JusticeÂ	371639	3018	250000000	2016	6.9	197000
13	Bryan Singer	434	169	0	18000	2E+08	Action|Adventure|Sci-Fi	Kevin Spacey	Superman ReturnsÂ	240396	2367	209000000	2006	6.1	0
14	Marc Forster	403	106	395	451	1.7E+08	Action|Adventure	Giancarlo Giannini	Quantum of SolaceÂ	330784	1243	200000000	2008	6.7	0
15	Gore Verbinski	313	151	563	40000	4.2E+08	Action|Adventure|Fantasy	Johnny Depp	Pirates of the Caribbean: Dead Man's ChestÂ	522040	1832	225000000	2006	7.3	5000
16	Gore Verbinski	450	150	563	40000	8.9E+07	Action|Adventure|Western	Johnny Depp	The Lone RangerÂ	181792	711	215000000	2013	6.5	48000
17	Zack Snyder	733	143	0	15000	2.9E+08	Action|Adventure|Fantasy|Sci-Fi	Henry Cavill	Man of SteelÂ	548573	2536	225000000	2013	7.2	118000
18	Andrew Adamson	258	150	80	22000	1.4E+08	Action|Adventure|Family|Fantasy	Peter Dinklage	The Chronicles of Narnia: Prince CaspianÂ	149922	438	225000000	2008	6.6	0
19	Joss Whedon	703	173	0	26000	6.2E+08	Action|Adventure|Sci-Fi	Chris Hemsworth	The AvengersÂ	995415	1722	220000000	2012	8.1	123000
20	Rob Marshall	448	136	252	40000	2.4E+08	Action|Adventure|Fantasy	Johnny Depp	Pirates of the Caribbean: On Stranger TidesÂ	370704	484	250000000	2011	6.7	58000
21	Barry Sonnenfeld	451	106	188	10000	1.8E+08	Action|Adventure|Comedy|Family|Fantasy|Sci-Fi	Will Smith	Men in Black 3Â	268154	341	225000000	2012	6.8	40000
22	Peter Jackson	422	164	0	5000	2.6E+08	Adventure|Fantasy	Aidan Turner	The Hobbit: The Battle of the Five ArmiesÂ	354228	802	250000000	2014	7.5	65000
23	Marc Webb	599	153	464	15000	2.6E+08	Action|Adventure|Fantasy	Emma Stone	The Amazing Spider-ManÂ	451803	1225	230000000	2012	7	56000
24	Ridley Scott	343	156	0	891	1.1E+08	Action|Adventure|Drama|History	Mark Addy	Robin HoodÂ	211765	546	200000000	2010	6.7	17000
25	Peter Jackson	509	186	0	5000	2.6E+08	Adventure|Fantasy	Aidan Turner	The Hobbit: The Desolation of SmaugÂ	483540	951	225000000	2013	7.9	83000
26	Chris Weitz	251	113	129	16000	7E+07	Adventure|Family|Fantasy	Christopher Lee	The Golden CompassÂ	149019	666	180000000	2007	6.1	0
27	Peter Jackson	446	201	0	6000	2.2E+08	Action|Adventure|Drama|Romance	Naomi Watts	King KongÂ	316018	2618	207000000	2005	7.2	0
28	James Cameron	315	194	0	29000	6.6E+08	Drama|Romance	Leonardo DiCaprio	TitanicÂ	793059	2528	200000000	1997	7.7	26000
29	Anthony Russo	516	147	94	21000	4.1E+08	Action|Adventure|Sci-Fi	Robert Downey Jr.	Captain America: Civil WarÂ	272670	1022	250000000	2016	8.2	72000
30	Peter Berg	377	131	532	14000	6.5E+07	Action|Adventure|Sci-Fi|Thriller	Liam Neeson	BattleshipÂ	202382	751	209000000	2012	5.9	44000
31	Colin Trevorrow	644	124	365	3000	6.5E+08	Action|Adventure|Sci-Fi|Thriller	Bryce Dallas Howard	Jurassic WorldÂ	418214	1290	150000000	2015	7	150000
32	Sam Mendes	750	143	0	883	3E+08	Action|Adventure|Thriller	Albert Finney	SkyfallÂ	522030	1498	200000000	2012	7.8	80000
33	Sam Raimi	300	135	0	24000	3.7E+08	Action|Adventure|Fantasy|Romance	J.K. Simmons	Spider-Man 2Â	411164	1303	200000000	2004	7.3	0
34	Shane Black	608	195	1000	21000	4.1E+08	Action|Adventure|Sci-Fi	Robert Downey Jr.	Iron Man 3Â	557489	1187	200000000	2013	7.2	95000
35	Tim Burton	451	108	13000	40000	3.3E+08	Adventure|Family|Fantasy	Johnny Depp	Alice in WonderlandÂ	306320	736	200000000	2010	6.5	24000
36	Brett Ratner	334	104	420	20000	2.3E+08	Action|Adventure|Fantasy|Sci-Fi|Thriller	Hugh Jackman	X-Men: The Last StandÂ	383427	1912	210000000	2006	6.8	0
37	Dan Scanlon	376	104	37	12000	2.7E+08	Adventure|Animation|Comedy|Family|Fantasy	Steve Buscemi	Monsters UniversityÂ	235025	265	200000000	2013	7.3	44000
38	Michael Bay	366	150	0	894	4E+08	Action|Adventure|Sci-Fi	Glenn Morshower	Transformers: Revenge of the FallenÂ	323207	1439	200000000	2009	6	0
39	Michael Bay	378	165	0	974	2.5E+08	Action|Adventure|Sci-Fi	Bingbing Li	Transformers: Age of ExtinctionÂ	242420	918	210000000	2014	5.7	56000
40	Sam Raimi	525	130	0	44000	2.3E+08	Adventure|Family|Fantasy	Tim Holmes	Oz the Great and PowerfulÂ	175409	511	215000000	2013	6.4	60000
41	Marc Webb	495	142	464	15000	2E+08	Action|Adventure|Fantasy|Sci-Fi	Emma Stone	The Amazing Spider-Man 2Â	321227	1067	200000000	2014	6.7	41000
42	Joseph Kosinski	469	125	364	12000	1.7E+08	Action|Adventure|Sci-Fi	Jeff Bridges	TRON: LegacyÂ	264183	665	170000000	2010	6.8	30000
43	John Lasseter	304	106	487	1000	1.9E+08	Adventure|Animation|Comedy|Family|Sport	Joe Mantegna	Cars 2Â	101178	283	200000000	2011	6.3	10000
44	Martin Campbell	436	123	258	16000	1.2E+08	Action|Adventure|Sci-Fi	Ryan Reynolds	Green LanternÂ	223393	550	200000000	2011	5.6	24000
45	Lee Unkrich	453	103	125	15000	4.1E+08	Adventure|Animation|Comedy|Family|Fantasy	Tom Hanks	Toy Story 3Â	544884	733	200000000	2010	8.3	30000
46	McG	422	118	368	23000	1.3E+08	Action|Adventure|Sci-Fi	Christian Bale	Terminator SalvationÂ	286095	974	200000000	2009	6.6	0
47	James Wan	424	140	0	26000	3.5E+08	Action|Crime|Thriller	Jason Statham	Furious 7Â	278232	657	190000000	2015	7.2	94000
48	Marc Forster	654	123	395	17000	2E+08	Action|Adventure|Horror|Sci-Fi|Thriller	Peter Capaldi	World War ZÂ	465019	995	190000000	2013	7	129000
49	Bryan Singer	539	149	0	34000	2.3E+08	Action|Adventure|Fantasy|Sci-Fi|Thriller	Jennifer Lawrence	X-Men: Days of Future PastÂ	514125	752	200000000	2014	8	82000
50	J.J. Abrams	590	132	14000	19000	2.3E+08	Action|Adventure|Sci-Fi	Benedict Cumberbatch	Star Trek Into DarknessÂ	395573	1171	190000000	2013	7.8	92000
51	Bryan Singer	338	114	0	979	6.5E+07	Adventure|Fantasy	Eddie Marsan	Jack the Giant SlayerÂ	106416	205	195000000	2013	6.3	22000
52	Baz Luhrmann	490	143	1000	29000	1.4E+08	Drama|Romance	Leonardo DiCaprio	The Great GatsbyÂ	362912	753	105000000	2013	7.3	115000
53	Mike Newell	306	116	179	15000	9.1E+07	Action|Adventure|Fantasy|Romance	Jake Gyllenhaal	Prince of Persia: The Sands of TimeÂ	222403	453	200000000	2010	6.6	23000
54	Guillermo del Toro	575	131	0	16000	1E+08	Action|Adventure|Sci-Fi	Charlie Hunnam	Pacific RimÂ	381148	1106	190000000	2013	7	83000
55	Michael Bay	428	154	0	894	3.5E+08	Action|Adventure|Sci-Fi	Glenn Morshower	Transformers: Dark of the MoonÂ	326180	899	195000000	2011	6.3	46000
56	Steven Spielberg	470	122	14000	11000	3.2E+08	Action|Adventure|Fantasy	Harrison Ford	Indiana Jones and the Kingdom of the Crystal SkullÂ	333847	2054	185000000	2008	6.2	5000

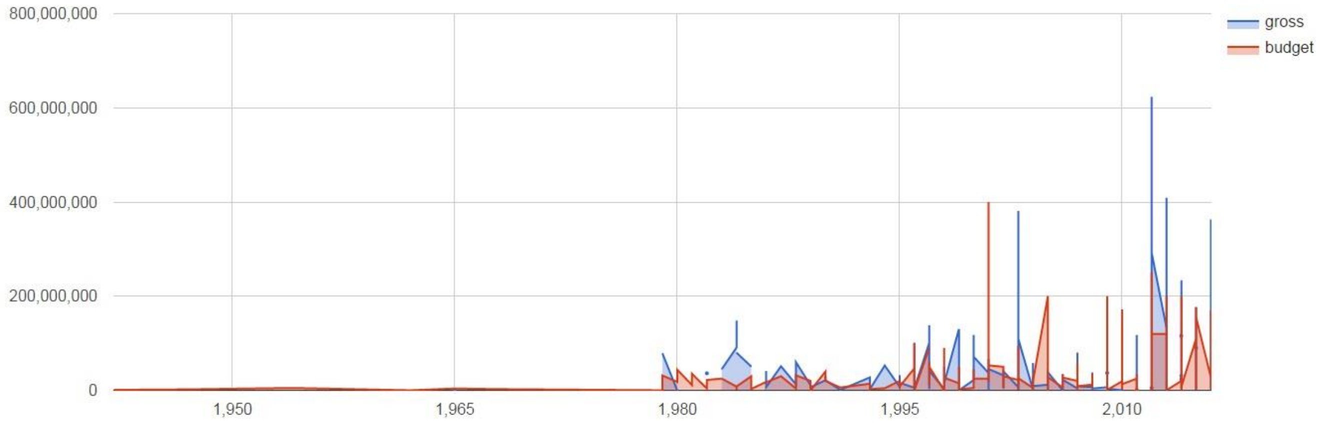
Stuff as links for the movie, second, third and fourth actor and their facebook likes has been removed

Description of Attributes

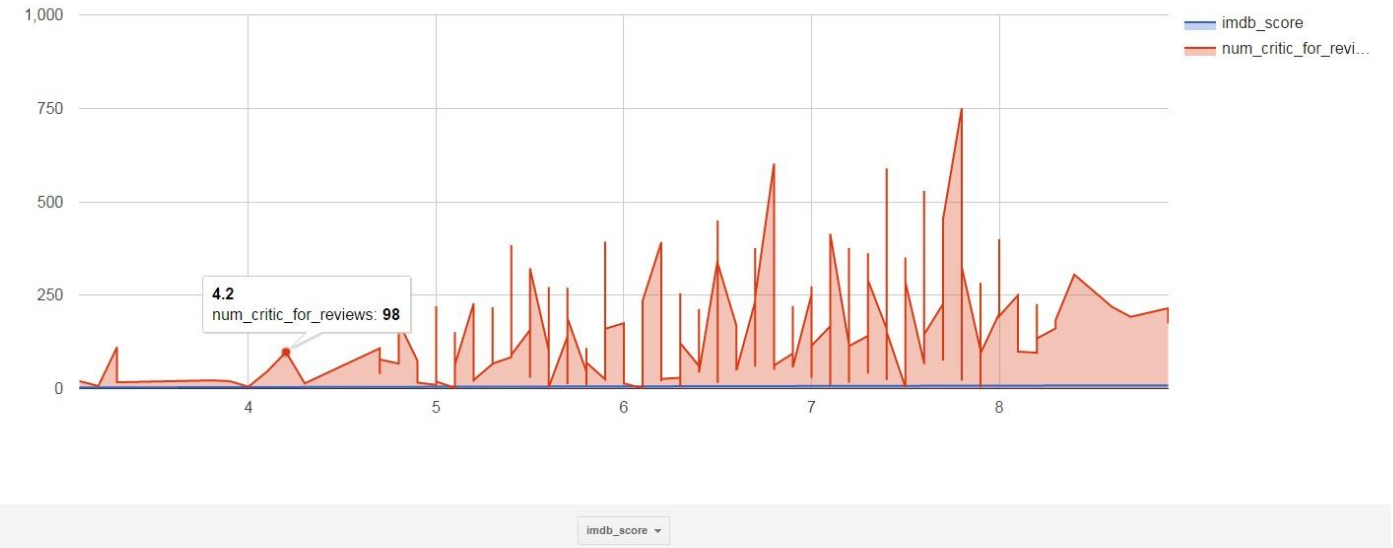
- Color – color of the movie (black and white or colored)
- Director – with more than 500 names
- Number of critic reviews
- Duration of the movies in minutes
- Director facebook likes
- Actor 1 , 2 , 3 names in different columns each
- Gross collected from the movie
- Genres – with a lot of different 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

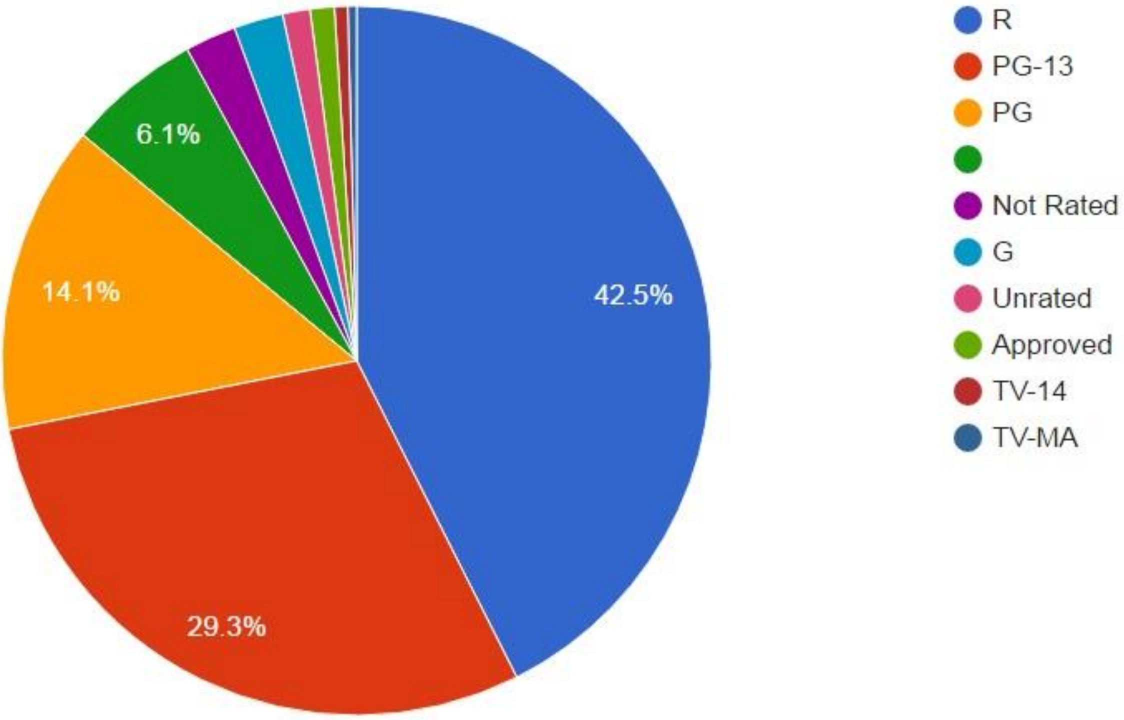
I decided to try and examine the most important part of the movies which is the amount of gross per budget gathered. As seen in the graph the early movies had a very small budget most ranging no more than a few millions for a big project and gross was not calculated very much back then. After year 1980 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).



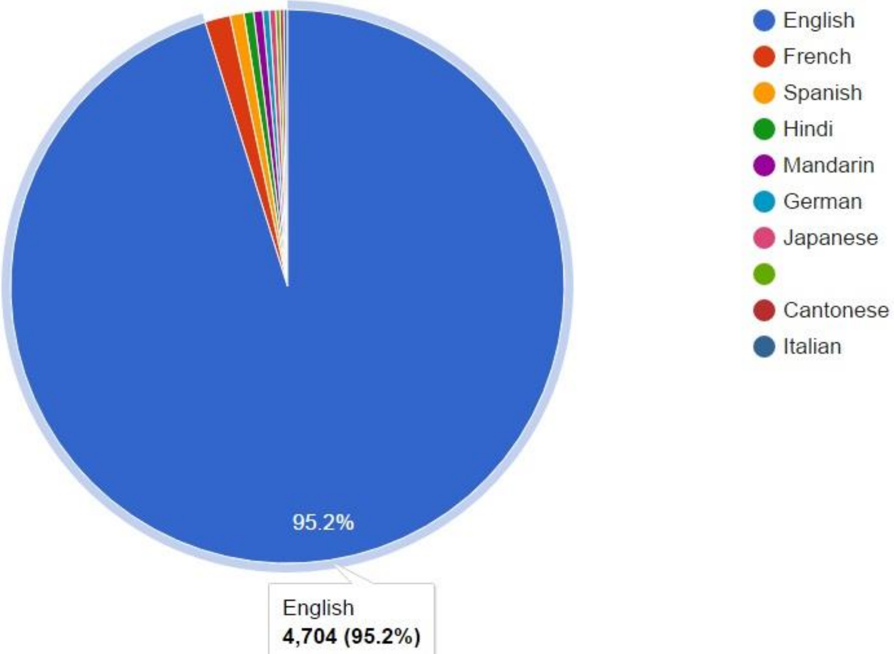
The graph shows the number of score from 0 to 10 and the amount of critic reviews given on them. As we can see the number is way too much for the movies which have higher than average rating and especially the ones above 7.5



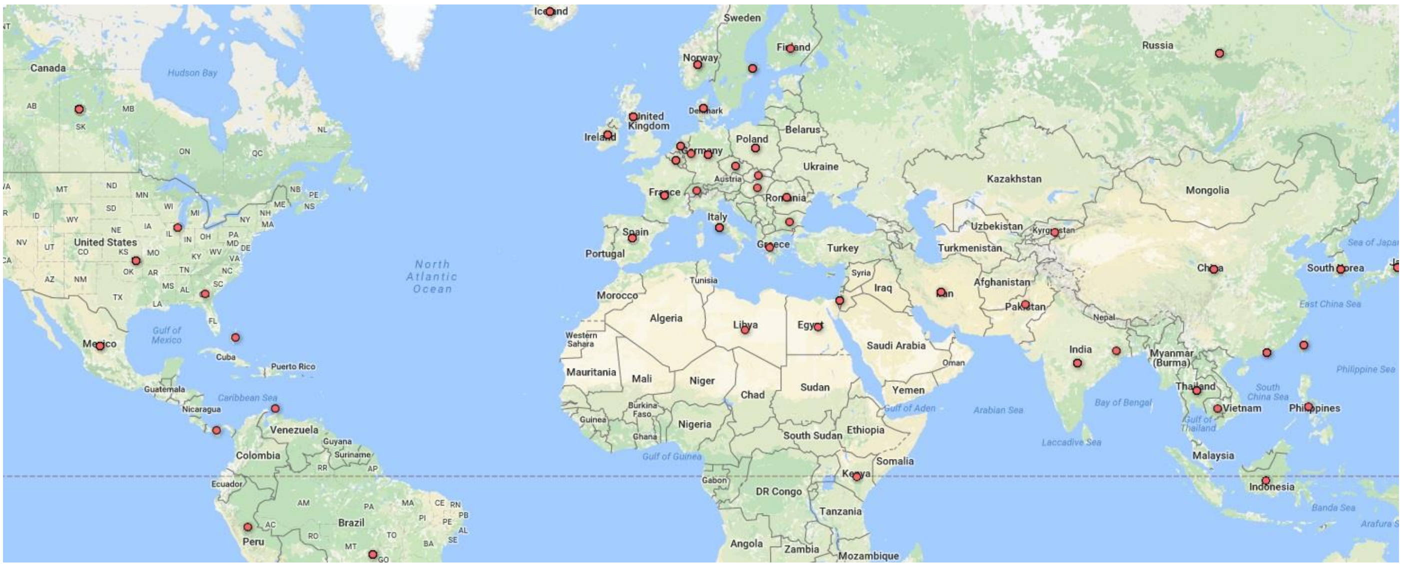
The rating of the gathered movies rated for different age of audience. The percent of “R” which stands for restricted type of movies has more than 40% which comes a little by surprise. Followed by more available to watch movies for kids from PG and PG13 with a total sum of 43%



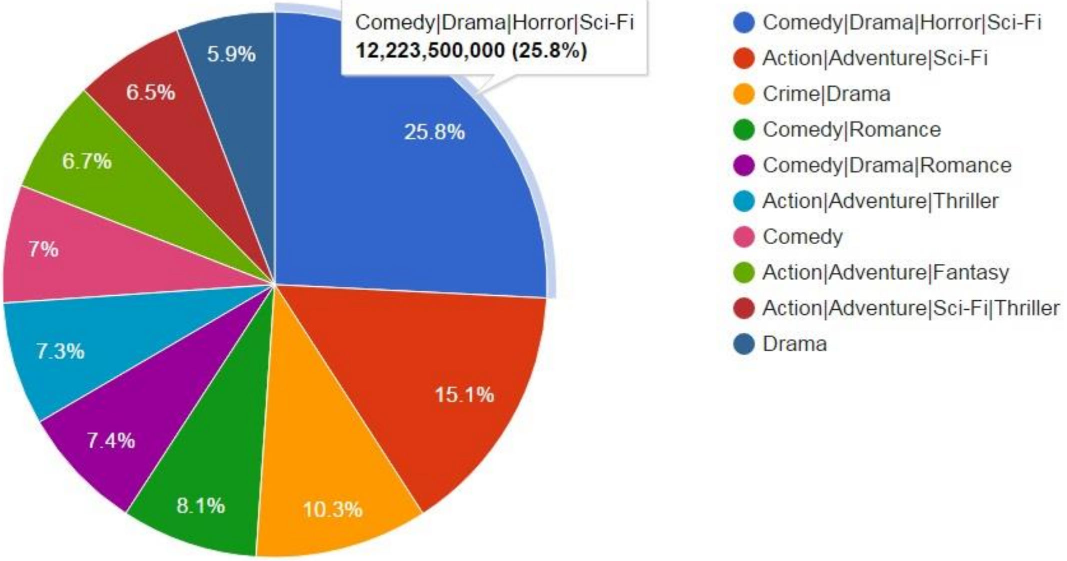
As most movies come from the United States with their huge industry and mostly in Hollywood with that percent raging to 95.2% as for sure that amount will be way less if all the international movies was included in the list.



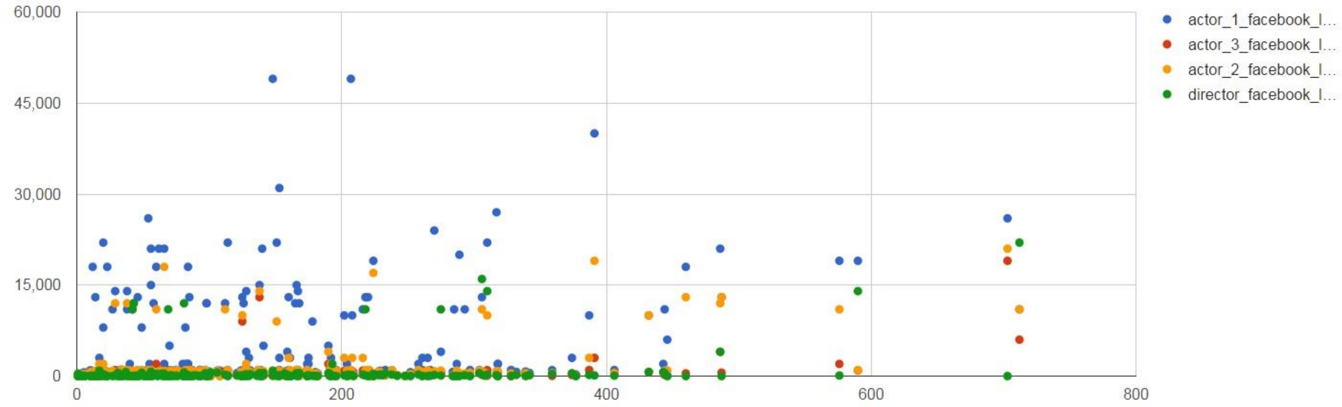
With the graph of languages was only appropriate to show the countries of origin and where the movies have been shot.



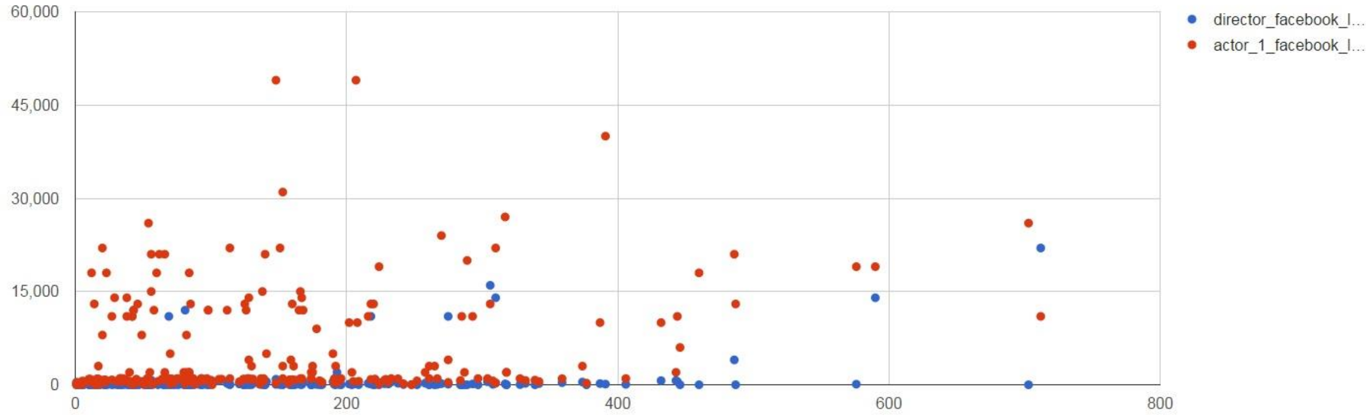
The data had a single problem, all the data had straight lines separating them rather than a comma which was really hard to visualize to either excel or google. But the data which was checked gave the result of the group of movies which had the tag of Comedy, Drama, Horror and Sci-Fi with an extraordinary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