Certification Program
on
Business Analytics

Assignment No:1

Date: 11th Feb to 17th Feb-2018

Submitted By:
Mallesham Yamulla

Topics

A. Make a brief Summary on the main theme of each paper.
(i) “Big data analytics”
(ii) “Big data and business analytics in a blended computing-business department”

B. Study the Health Statistics Dataset and formulate the different classes of Business Analytics Problems (from the discussed 9 Analytics Prospective from Business)

C. Perform basic Data Analysis (Data Types, Graphical & Descriptive)on the attached “Health Statistics Dataset- Revised” (which includes the latest data points of your batch) and give your findings.

D. Read and understand the concept behind Kano Model & QFD in Business Problem Formulation and prepare a brief write-up. (preferably identifying a business scenario from your own organisation)

Big Data Analytics

Introduction to Big Data Analytics

Big Data: As days pass by, Organisations increasingly need to analyse information to make decisions for achieving greater efficiency, profits and productivity. As relational databases have grown in size to satisfy these requirements, organisations have also looked at other technologies for storing vast amounts of information, these new systems are often referred to under the umbrella term BIG DATA.

Here, To be big data, it needs to have all three of these below attributes,

1) Do we have a high **volume** of data?

- If we are collecting petabytes of data each day we probably have enough volume, in the near future maybe exabyte will be considered a high enough volume.

2) Do we have a wide **variety** of data?

- There should be a wide variety of information, there can be text, videos, sound and pictures.

3) Is the data coming in at a high **velocity**?

- Let's think of Stock exchange, here they handle billions of transactions each day, the stock prices are pouring in and fluctuating in milliseconds and they have a high volume of data coming in at a high velocity.

Big Data Analytics

Big Data Analytics:

Big data analytics leverages distributed computing technologies and data analytics techniques to overcome computational challenges presented by big data sets.

Distributed computing means an approach used in computer science to break down a task into smaller pieces that are easier to get processed, for example a fragment of my task can be processed in India while another piece can be worked in Mexico.

Could computing provides a platform on which distributed computing can be implemented with low cost and scalable methods.

To simply put in, cloud computing offers a bunch of computers housed in data center, in addition to the hardware a software solution is necessary to manage various aspects of distributed computing, this is why we need software tools such as Hadoop and NoSQL databases, once we get with these software and hardware infrastructure to store, manage and process the big data sets we are finally ready to carry out data analytics methods.

Here Data Analytics is the process of transforming data into insight for making better decisions and it is used for data driven or fact-based decision making which is often seen as more objective than other alternatives for decision making.

Type of Data Analytics:

1. Descriptive Analytics: What has happened ??
2. Predictive Analytics: What might happen ??
3. Prescriptive Analytics: What should we do ??

Big Data Analytics

Why should we put Big data and Analytics together now ?

1. Big data can be handled with the various analytical tools and databases- In the recent days the different kind of analytical tools and data storages have been made available in the market to analyse the big data such as R, Python, Spark, Scala, NoSQL databases like MongoDB, Hive and Cassandra etc..

2. We can learn a lot from messy data as long as it's big. We often deal with many different types of data in the analytics, and this data can be available in.

    A. **Structured:** Data that follows a specific format in a specific order

    B. **Semi- structured:** Data with some structure, but also with added flexibility to change field names and create values

    C. **Unstructured:** Data that doesn't follow a schema and has no data model.

    The more messier data we have the more better insights we would find out.

3. Affordable costs on executing analytics projects.- There are Cloud platforms available on which the analytics projects can be implemented and maintained at reasonable prices like AWS, Google and EMC etc etc.

Big Data Analytics

Big data Problem or Opportunity ?

1. Big data would be mostly an opportunity not a problem.

   We are into telecom, we have a system called NPP(Network Protection Platform) that detects the spams across the globe, here few years ago we used to store the messaging information in the traditional systems where the data retention period would be 1week, we have come under the threat intelligence department to carry out an analysis of this stored information to get the stats of

   How much of spams have passed through from the networks daily, weekly and Monthly?

   How was the subscribers behaviours, which country the spam was coming from and which country the spam was going to?

   How were the messaging patterns? And So on and so forth.

   We have found it difficult to get the above questions answered by analysing the limited information of 1 week in systems and Here in this situation the big data has come into picture to help us out with larger storage, now in our systems the data retention period has got extended to 3-6months with the help of Hadoop system.

   We have already started doing data analytics on this larger data, we have come through dealing with Spams in out networks at higher rates and the blocking rate of spams has been going up moderately from day to day and it is estimated that there has been an improvement of 75% in our works with the help of Big Data.

Big Data Analytics

Tools, Techniques and Trends for Big Data Analytics

Big data technologies have wide and long list of their applications. It is used for Search Engine, Log Processing, Recommender System, Data Warehousing, Video and Image Analysis, Banking & Financial, Telecom, Retail, Manufacturing, Web & Social Media, Medicine, Healthcare, Science & Research and Social Life.

There are many options for big data analytics tools, however it's hard know them all and select the best one and the tools needed by data analytics team fall into three general categories;

A. **Software to hold the data:** These re the spreadsheets, databases, and key/value stores.Some papular software includes Hadoop, Cassandra, and PostgreSQL.

B. **Tools used to scrub the data:** Data scrubbing, also called cleansing, makes data easier to work with by modifying or amending data or deleting duplicate, incorrectly formatted ,incorrect ,or incomplete data.Typical tools used scrub data are text editors, scripting tools, and programming languages such as Python and Scala.

C. **Statistical packages to help analyse the data:** The most popular are the open source software environment R,IBM SPSS, Predictive analytics software ,and Pythons programming language.Most of these include the ability to visualise the data.you will need this to make nice charts and graphs.

Big Data Analytics

Best practices for Business analytics

A. Creating dimensions of all the data being store is a good practice for Big data analytics. It needs to be divided into dimensions and facts.

B. All the dimensions should have durable surrogate keys meaning that these keys can’t be changed by any business rule and are assigned in sequence or generated by some hashing algorithm ensuring uniqueness.

C. Expect to integrate structured and unstructured data as all kind of data is a part of Big data which needs to be analysed together.

D. Generality of the technology is needed to deal with different formats of data. Building technology around key value pairs work.

E. Analysing data sets including identifying information about individuals or organisations privacy is an issue whose importance particularly to consumers is growing as the value of Big data becomes more apparent.

F. Data quality needs to be better. Different tasks like filtering, cleansing, pruning, conforming, matching, joining, and diagnosing should be applied at the earliest touch points possible.

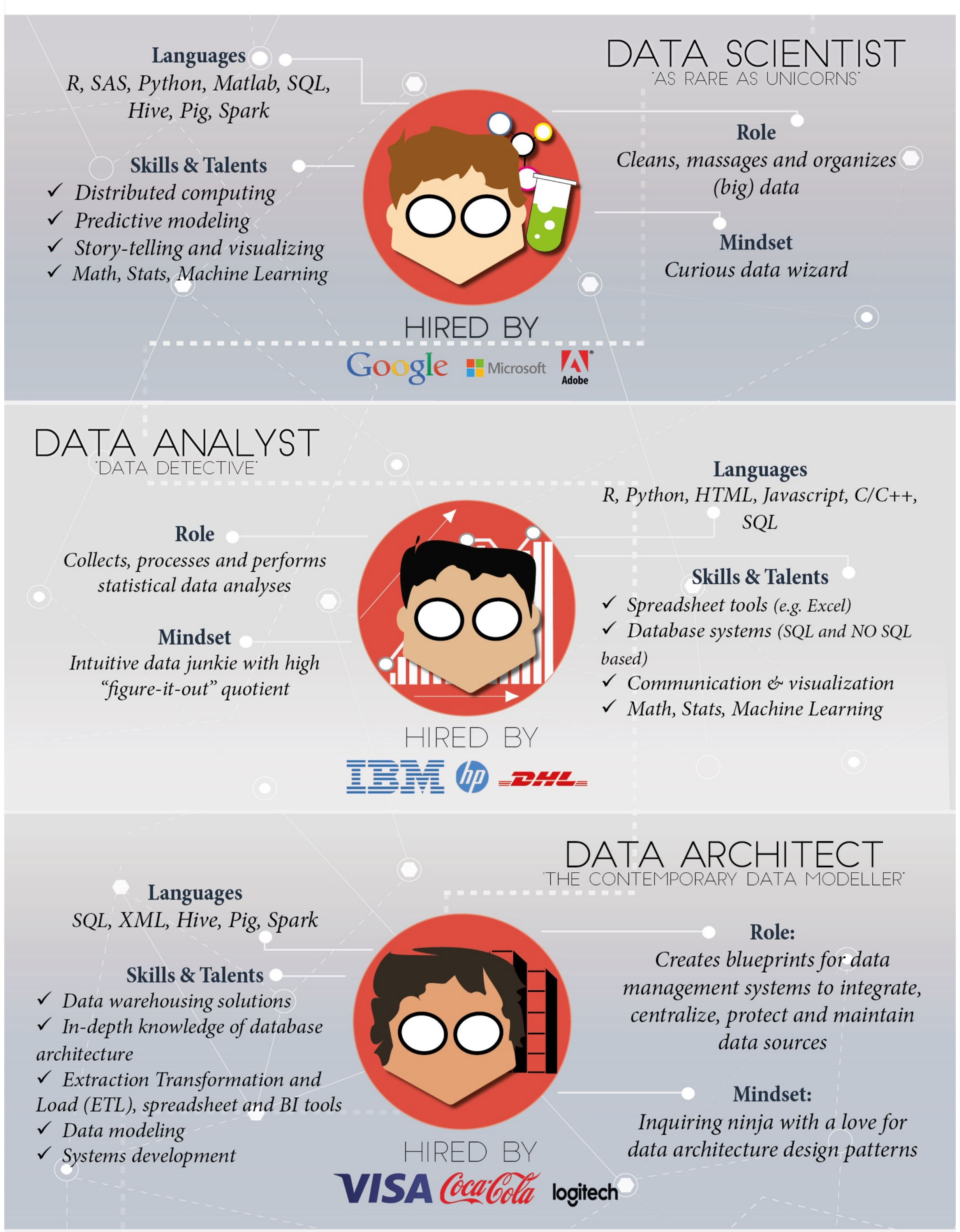
G. There should be certain limits on the scalability of the data stored.

H. Business leaders and IT leaders should work together to yield more business value from the data. Collecting, storing and analysing data comes at a cost. Business leaders will go for it but IT leaders have to look for many things like technological limitations, staff restrictions etc. The decisions taken should be revised to ensure that the organisation is considering the right data to produce insights at any given point of time.

I. Investment in data quality and metadata is also important as it reduces the processing time

Big Data and Business Analytics in a Blended Computing - Business Department

Building analytics Team



DATA SCIENTIST
'AS RARE AS UNICORNS'

Languages
R, SAS, Python, Matlab, SQL, Hive, Pig, Spark

Skills & Talents
- ✓ Distributed computing
- ✓ Predictive modeling
- ✓ Story-telling and visualizing
- ✓ Math, Stats, Machine Learning

Role
Cleans, massages and organizes (big) data

Mindset
Curious data wizard

HIRED BY
Google Microsoft Adobe

DATA ANALYST
'DATA DETECTIVE'

Role
Collects, processes and performs statistical data analyses

Mindset
Intuitive data junkie with high "figure-it-out" quotient

Languages
R, Python, HTML, Javascript, C/C++, SQL

Skills & Talents
- ✓ Spreadsheet tools (e.g. Excel)
- ✓ Database systems (SQL and NO SQL based)
- ✓ Communication & visualization
- ✓ Math, Stats, Machine Learning

HIRED BY
IBM hp DHL

DATA ARCHITECT
'THE CONTEMPORARY DATA MODELLER'

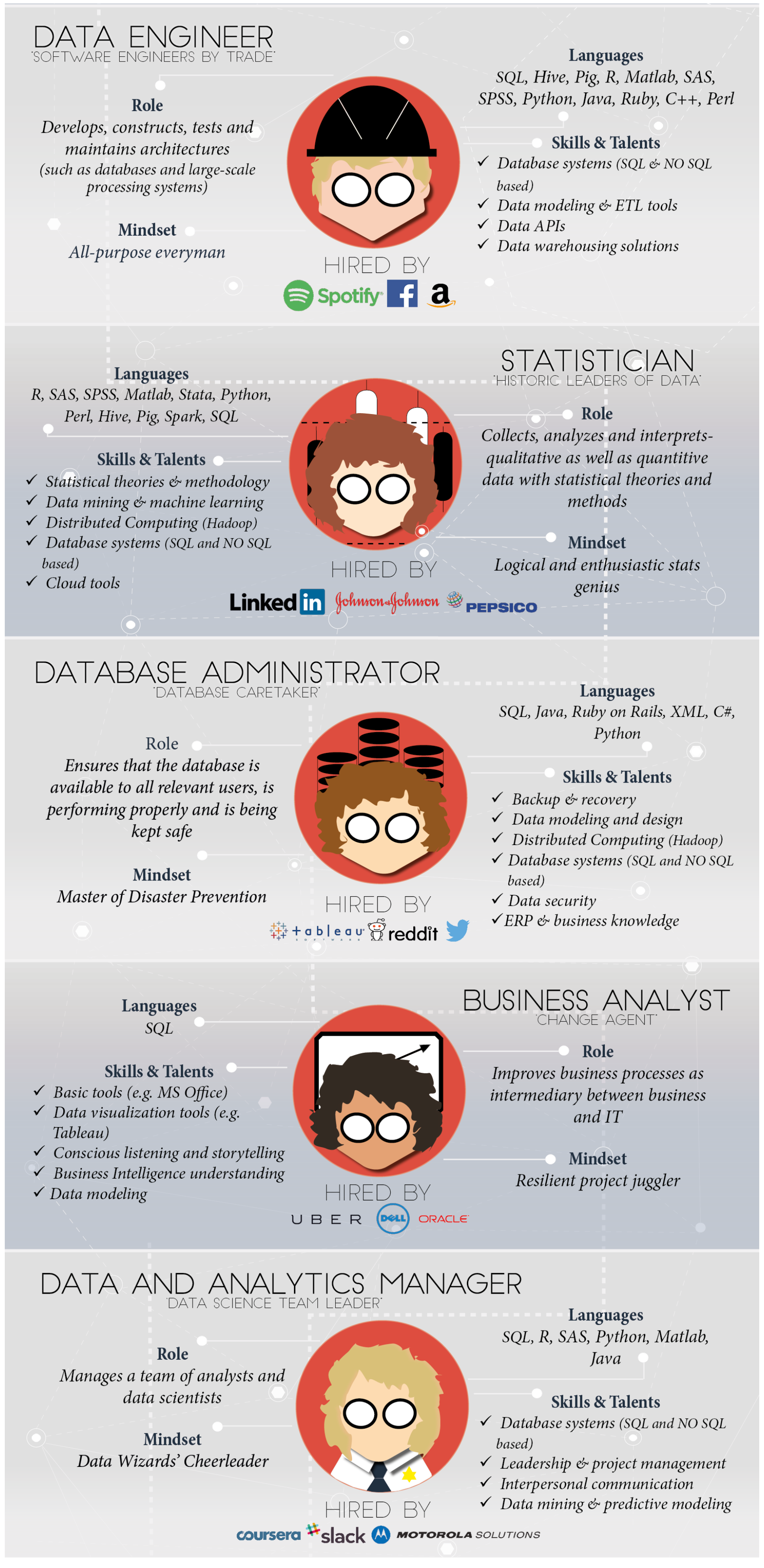
Languages
SQL, XML, Hive, Pig, Spark

Skills & Talents
- ✓ Data warehousing solutions
- ✓ In-depth knowledge of database architecture
- ✓ Extraction Transformation and Load (ETL), spreadsheet and BI tools
- ✓ Data modeling
- ✓ Systems development

Role:
Creates blueprints for data management systems to integrate, centralize, protect and maintain data sources

Mindset:
Inquiring ninja with a love for data architecture design patterns

HIRED BY
VISA Coca-Cola logitech



DATA ENGINEER
'SOFTWARE ENGINEERS BY TRADE'

Role
Develops, constructs, tests and maintains architectures (such as databases and large-scale processing systems)

Mindset
All-purpose everyman

Languages
SQL, Hive, Pig, R, Matlab, SAS, SPSS, Python, Java, Ruby, C++, Perl

Skills & Talents
- ✓ Database systems (SQL & NO SQL based)
- ✓ Data modeling & ETL tools
- ✓ Data APIs
- ✓ Data warehousing solutions

HIRED BY
Spotify f a

STATISTICIAN
'HISTORIC LEADERS OF DATA'

Languages
R, SAS, SPSS, Matlab, Stata, Python, Perl, Hive, Pig, Spark, SQL

Skills & Talents
- ✓ Statistical theories & methodology
- ✓ Data mining & machine learning
- ✓ Distributed Computing (Hadoop)
- ✓ Database systems (SQL and NO SQL based)
- ✓ Cloud tools

Role
Collects, analyzes and interprets-qualitative as well as quantitive data with statistical theories and methods

Mindset
Logical and enthusiastic stats genius

HIRED BY
LinkedIn Johnson&Johnson PEPSICO

DATABASE ADMINISTRATOR
'DATABASE CARETAKER'

Role
Ensures that the database is available to all relevant users, is performing properly and is being kept safe

Mindset
Master of Disaster Prevention

Languages
SQL, Java, Ruby on Rails, XML, C#, Python

Skills & Talents
- ✓ Backup & recovery
- ✓ Data modeling and design
- ✓ Distributed Computing (Hadoop)
- ✓ Database systems (SQL and NO SQL based)
- ✓ Data security
- ✓ ERP & business knowledge

HIRED BY
tableau reddit

BUSINESS ANALYST
'CHANGE AGENT'

Languages
SQL

Skills & Talents
- ✓ Basic tools (e.g. MS Office)
- ✓ Data visualization tools (e.g. Tableau)
- ✓ Conscious listening and storytelling
- ✓ Business Intelligence understanding
- ✓ Data modeling

Role
Improves business processes as intermediary between business and IT

Mindset
Resilient project juggler

HIRED BY
UBER DELL ORACLE

DATA AND ANALYTICS MANAGER
'DATA SCIENCE TEAM LEADER'

Role
Manages a team of analysts and data scientists

Mindset
Data Wizards' Cheerleader

Languages
SQL, R, SAS, Python, Matlab, Java

Skills & Talents
- ✓ Database systems (SQL and NO SQL based)
- ✓ Leadership & project management
- ✓ Interpersonal communication
- ✓ Data mining & predictive modeling

HIRED BY
coursera slack MOTOROLA SOLUTIONS

Rounding out our talent:

1. Does the candidate have solid programming skills?

2. Does the candidate excel at producing analytics for computers or humans? (And which do you need?)

3. Can the candidate provide concrete examples of when she has improved a business process through her work?

4. Is the candidate a good communicator?

5. Can the candidate be creative and open-minded?

6. Does the candidate have solid business understanding?

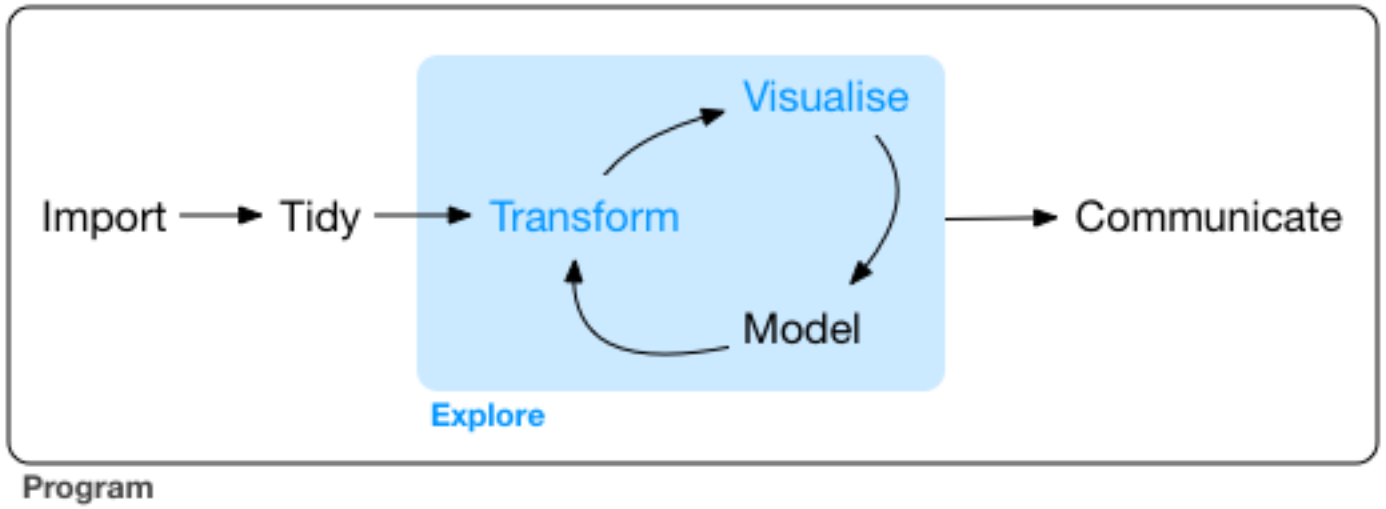
Exploratory Data Analysis of Health DataSet

Exploratory Data Analysis (EDA) is used on the one hand to answer questions, test business assumptions, generate hypotheses for further analysis. On the other hand, we can also use it to prepare the data for modeling. The thing that these two probably have in common is a good knowledge of our data to either get the answers that we need or to develop an intuition for interpreting the results of future modeling.

There are a lot of ways to reach these goals: we can get a basic description of the data, visualise it, identify patterns in it, identify challenges of using the data, etc.

One of the things that we will often see when w're reading about EDA is Data profiling. Data profiling is concerned with summarising our dataset through descriptive statistics. We want to use a variety of measurements to better understand our dataset.

The goal of data profiling is to have a solid understanding of our data so we can afterwards start querying and visualising our data in various ways. However, this doesn't mean that we don't have to iterate: exactly because data profiling is concerned with summarising our dataset, it is frequently used to assess the data quality. Depending on the result of the data profiling, we might decide to correct, discard or handle our data differently



Exploratory Data Analysis of Health DataSet

Data Exploration Steps:

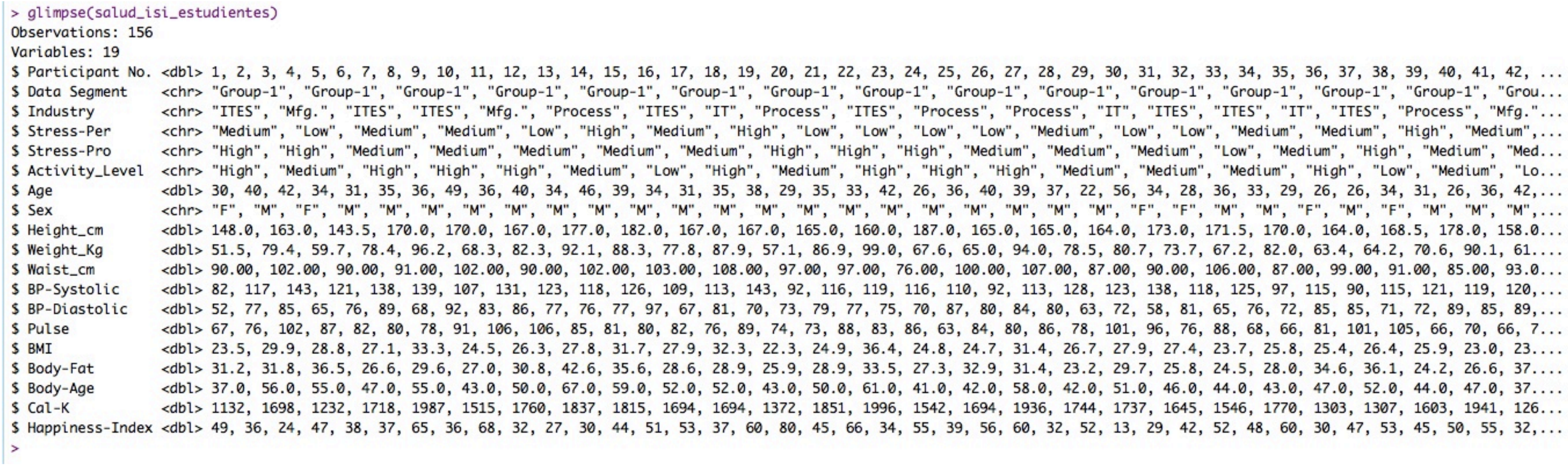
EDA of health data has been executed by R programming with the help of it's packages like Tidyverse, here The tidyverse is a set of packages that work in harmony because they share common data representations and API design. The **tidyverse** package is designed to make it easy to install and load core packages from the tidyverse in a single command, it has contained the below given set of packages.

Package	Purpose
Ggplot2	For data visualisation
Dplyr	For data transformations
Tidyr	For tidying
Readr	For data importing
Purrr	For functional programming
Tibble	For tibbles, a modern re-imagining of data frames

Exploratory Data Analysis of Health DataSet

Data Exploration Steps:

Importing : The health data set is available in .xlsx format, it has got 2 sheets and One sheet contains all the groups students info and another one has the latest group students info only, so I have used a package called readxl in R to get the sheet-1 data loaded into R as Data frame to do EDA on it and it's named as salud_isi_estudiantes.

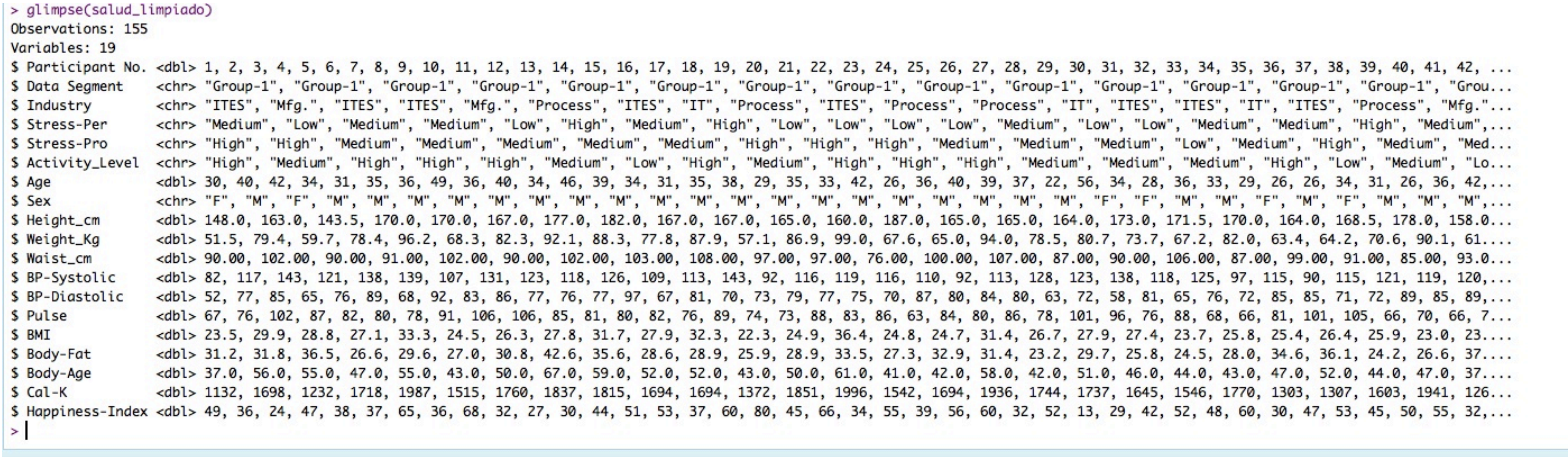


```
> glimpse(salud_isi_estudientes)
Observations: 156
Variables: 19
$ Participant No. <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, ...
$ Data Segment    <chr> "Group-1", "Group-1", "Group-1", "Group-1", "Group-1", "Group-1", "Group-1", "Group-1", "Group-1", "Group-1", "Group-1", "Group-1", "Group-1", "Group-1", "Grou...
$ Industry        <chr> "ITES", "Mfg.", "ITES", "ITES", "Mfg.", "Process", "ITES", "IT", "Process", "ITES", "Process", "Process", "IT", "ITES", "ITES", "IT", "ITES", "Process", "Mfg."...
$ Stress-Per      <chr> "Medium", "Low", "Medium", "Medium", "Low", "High", "Medium", "High", "Low", "Low", "Low", "Low", "Medium", "Low", "Low", "Medium", "Medium", "High", "Medium",...
$ Stress-Pro      <chr> "High", "High", "Medium", "Medium", "Medium", "Medium", "Medium", "High", "High", "High", "Medium", "Medium", "Medium", "Low", "Medium", "High", "Medium", "Med...
$ Activity_Level  <chr> "High", "Medium", "High", "High", "High", "Medium", "Low", "High", "Medium", "High", "High", "High", "Medium", "Medium", "Medium", "High", "Low", "Medium", "Lo...
$ Age             <dbl> 30, 40, 42, 34, 31, 35, 36, 49, 36, 40, 34, 46, 39, 34, 31, 35, 38, 29, 35, 33, 42, 26, 36, 40, 39, 37, 22, 56, 34, 28, 36, 33, 29, 26, 26, 34, 31, 26, 36, 42,...
$ Sex             <chr> "F", "M", "F", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "F", "F", "M", "M", "F", "M", "F", "M", "M", "M",...
$ Height_cm       <dbl> 148.0, 163.0, 143.5, 170.0, 170.0, 167.0, 177.0, 182.0, 167.0, 167.0, 165.0, 160.0, 187.0, 165.0, 165.0, 164.0, 173.0, 171.5, 170.0, 164.0, 168.5, 178.0, 158.0...
$ Weight_Kg       <dbl> 51.5, 79.4, 59.7, 78.4, 96.2, 68.3, 82.3, 92.1, 88.3, 77.8, 87.9, 57.1, 86.9, 99.0, 67.6, 65.0, 94.0, 78.5, 80.7, 73.7, 67.2, 82.0, 63.4, 64.2, 70.6, 90.1, 61....
$ Waist_cm        <dbl> 90.00, 102.00, 90.00, 91.00, 102.00, 90.00, 102.00, 103.00, 108.00, 97.00, 97.00, 76.00, 100.00, 107.00, 87.00, 90.00, 106.00, 87.00, 99.00, 91.00, 85.00, 93.0...
$ BP-Systolic     <dbl> 82, 117, 143, 121, 138, 139, 107, 131, 123, 118, 126, 109, 113, 143, 92, 116, 119, 116, 110, 92, 113, 128, 123, 138, 118, 125, 97, 115, 90, 115, 121, 119, 120,...
$ BP-Diastolic    <dbl> 52, 77, 85, 65, 76, 89, 68, 92, 83, 86, 77, 76, 77, 97, 67, 81, 70, 73, 79, 77, 75, 70, 87, 80, 84, 80, 63, 72, 58, 81, 65, 76, 72, 85, 85, 71, 72, 89, 85, 89,...
$ Pulse           <dbl> 67, 76, 102, 87, 82, 80, 78, 91, 106, 106, 85, 81, 80, 82, 76, 89, 74, 73, 88, 83, 86, 63, 84, 80, 86, 78, 101, 96, 76, 88, 68, 66, 81, 101, 105, 66, 70, 66, 7...
$ BMI             <dbl> 23.5, 29.9, 28.8, 27.1, 33.3, 24.5, 26.3, 27.8, 31.7, 27.9, 32.3, 22.3, 24.9, 36.4, 24.8, 24.7, 31.4, 26.7, 27.9, 27.4, 23.7, 25.8, 25.4, 26.4, 25.9, 23.0, 23....
$ Body-Fat        <dbl> 31.2, 31.8, 36.5, 26.6, 29.6, 27.0, 30.8, 42.6, 35.6, 28.6, 28.9, 25.9, 28.9, 33.5, 27.3, 32.9, 31.4, 23.2, 29.7, 25.8, 24.5, 28.0, 34.6, 36.1, 24.2, 26.6, 37....
$ Body-Age        <dbl> 37.0, 56.0, 55.0, 47.0, 55.0, 43.0, 50.0, 67.0, 59.0, 52.0, 52.0, 43.0, 50.0, 61.0, 41.0, 42.0, 58.0, 42.0, 51.0, 46.0, 44.0, 43.0, 47.0, 52.0, 44.0, 47.0, 37....
$ Cal-K           <dbl> 1132, 1698, 1232, 1718, 1987, 1515, 1760, 1837, 1815, 1694, 1694, 1372, 1851, 1996, 1542, 1694, 1936, 1744, 1737, 1645, 1546, 1770, 1303, 1307, 1603, 1941, 126...
$ Happiness-Index <dbl> 49, 36, 24, 47, 38, 37, 65, 36, 68, 32, 27, 30, 44, 51, 53, 37, 60, 80, 45, 66, 34, 55, 39, 56, 60, 32, 52, 13, 29, 42, 52, 48, 60, 30, 47, 53, 45, 50, 55, 32,...
>
```

Exploratory Data Analysis of Health DataSet

Data Exploration Steps:

Cleaning : The available data frame salud_isi_estudiantes is found to be having got the clean data, however there are missing values in one observation out of 156, it has been taken out form this data frame and formed a new data frame called salud_limpiado to get clear-cut insight from it.



```
> glimpse(salud_limpiado)
Observations: 155
Variables: 19
$ Participant No. <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, ...
$ Data Segment    <chr> "Group-1", "Group-1", "Group-1", "Group-1", "Group-1", "Group-1", "Group-1", "Group-1", "Group-1", "Group-1", "Group-1", "Group-1", "Group-1", "Group-1", "Grou...
$ Industry        <chr> "ITES", "Mfg.", "ITES", "ITES", "Mfg.", "Process", "ITES", "IT", "Process", "ITES", "Process", "Process", "IT", "ITES", "ITES", "IT", "ITES", "Process", "Mfg."...
$ Stress-Per      <chr> "Medium", "Low", "Medium", "Medium", "Low", "High", "Medium", "High", "Low", "Low", "Low", "Low", "Medium", "Low", "Low", "Medium", "Medium", "High", "Medium",...
$ Stress-Pro      <chr> "High", "High", "Medium", "Medium", "Medium", "Medium", "Medium", "High", "High", "High", "Medium", "Medium", "Medium", "Low", "Medium", "High", "Medium", "Med...
$ Activity_Level  <chr> "High", "Medium", "High", "High", "High", "Medium", "Low", "High", "Medium", "High", "High", "High", "Medium", "Medium", "Medium", "High", "Low", "Medium", "Lo...
$ Age             <dbl> 30, 40, 42, 34, 31, 35, 36, 49, 36, 40, 34, 46, 39, 34, 31, 35, 38, 29, 35, 33, 42, 26, 36, 40, 39, 37, 22, 56, 34, 28, 36, 33, 29, 26, 26, 34, 31, 26, 36, 42,...
$ Sex             <chr> "F", "M", "F", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "F", "F", "M", "M", "F", "M", "F", "M", "M", "M",...
$ Height_cm       <dbl> 148.0, 163.0, 143.5, 170.0, 170.0, 167.0, 177.0, 182.0, 167.0, 167.0, 165.0, 160.0, 187.0, 165.0, 165.0, 164.0, 173.0, 171.5, 170.0, 164.0, 168.5, 178.0, 158.0...
$ Weight_Kg       <dbl> 51.5, 79.4, 59.7, 78.4, 96.2, 68.3, 82.3, 92.1, 88.3, 77.8, 87.9, 57.1, 86.9, 99.0, 67.6, 65.0, 94.0, 78.5, 80.7, 73.7, 67.2, 82.0, 63.4, 64.2, 70.6, 90.1, 61....
$ Waist_cm        <dbl> 90.00, 102.00, 90.00, 91.00, 102.00, 90.00, 102.00, 103.00, 108.00, 97.00, 97.00, 76.00, 100.00, 107.00, 87.00, 90.00, 106.00, 87.00, 99.00, 91.00, 85.00, 93.0...
$ BP-Systolic     <dbl> 82, 117, 143, 121, 138, 139, 107, 131, 123, 118, 126, 109, 113, 143, 92, 116, 119, 116, 110, 92, 113, 128, 123, 138, 118, 125, 97, 115, 90, 115, 121, 119, 120,...
$ BP-Diastolic    <dbl> 52, 77, 85, 65, 76, 89, 68, 92, 83, 86, 77, 76, 77, 97, 67, 81, 70, 73, 79, 77, 75, 70, 87, 80, 84, 80, 63, 72, 58, 81, 65, 76, 72, 85, 85, 71, 72, 89, 85, 89,...
$ Pulse           <dbl> 67, 76, 102, 87, 82, 80, 78, 91, 106, 106, 85, 81, 80, 82, 76, 89, 74, 73, 88, 83, 86, 63, 84, 80, 86, 78, 101, 96, 76, 88, 68, 66, 81, 101, 105, 66, 70, 66, 7...
$ BMI             <dbl> 23.5, 29.9, 28.8, 27.1, 33.3, 24.5, 26.3, 27.8, 31.7, 27.9, 32.3, 22.3, 24.9, 36.4, 24.8, 24.7, 31.4, 26.7, 27.9, 27.4, 23.7, 25.8, 25.4, 26.4, 25.9, 23.0, 23....
$ Body-Fat        <dbl> 31.2, 31.8, 36.5, 26.6, 29.6, 27.0, 30.8, 42.6, 35.6, 28.6, 28.9, 25.9, 28.9, 33.5, 27.3, 32.9, 31.4, 23.2, 29.7, 25.8, 24.5, 28.0, 34.6, 36.1, 24.2, 26.6, 37....
$ Body-Age        <dbl> 37.0, 56.0, 55.0, 47.0, 55.0, 43.0, 50.0, 67.0, 59.0, 52.0, 52.0, 43.0, 50.0, 61.0, 41.0, 42.0, 58.0, 42.0, 51.0, 46.0, 44.0, 43.0, 47.0, 52.0, 44.0, 47.0, 37....
$ Cal-K           <dbl> 1132, 1698, 1232, 1718, 1987, 1515, 1760, 1837, 1815, 1694, 1694, 1372, 1851, 1996, 1542, 1694, 1936, 1744, 1737, 1645, 1546, 1770, 1303, 1307, 1603, 1941, 126...
$ Happiness-Index <dbl> 49, 36, 24, 47, 38, 37, 65, 36, 68, 32, 27, 30, 44, 51, 53, 37, 60, 80, 45, 66, 34, 55, 39, 56, 60, 32, 52, 13, 29, 42, 52, 48, 60, 30, 47, 53, 45, 50, 55, 32,...
>
```

Exploratory Data Analysis of Health DataSet

Data Exploration Steps:

Transforming : As part of data transformation method I would like add up the below 3 new variables to the data frame to see out and categorise the group of students based on the available information of their BP-Systolic, BP-Diastolic, BMI Levels and Pulse rates.

Blood pressure levels:

BP-Systolic	BP-Diastolic	Blood_pressure_levels
<120	<80	Normal
Between(120,139)	Between(80,89)	Prehyper
Between(120,139)	Between(120,139)	Stage 1 hypertension
>=160	>=100	Stage 2 hypersion

Weight levels:

BMI Level	Weight_levels
<18.5	Underweight
Between(18.5,24.9)	Normal Weight
Between(25,29.5)	Over weight
>=30	Obese

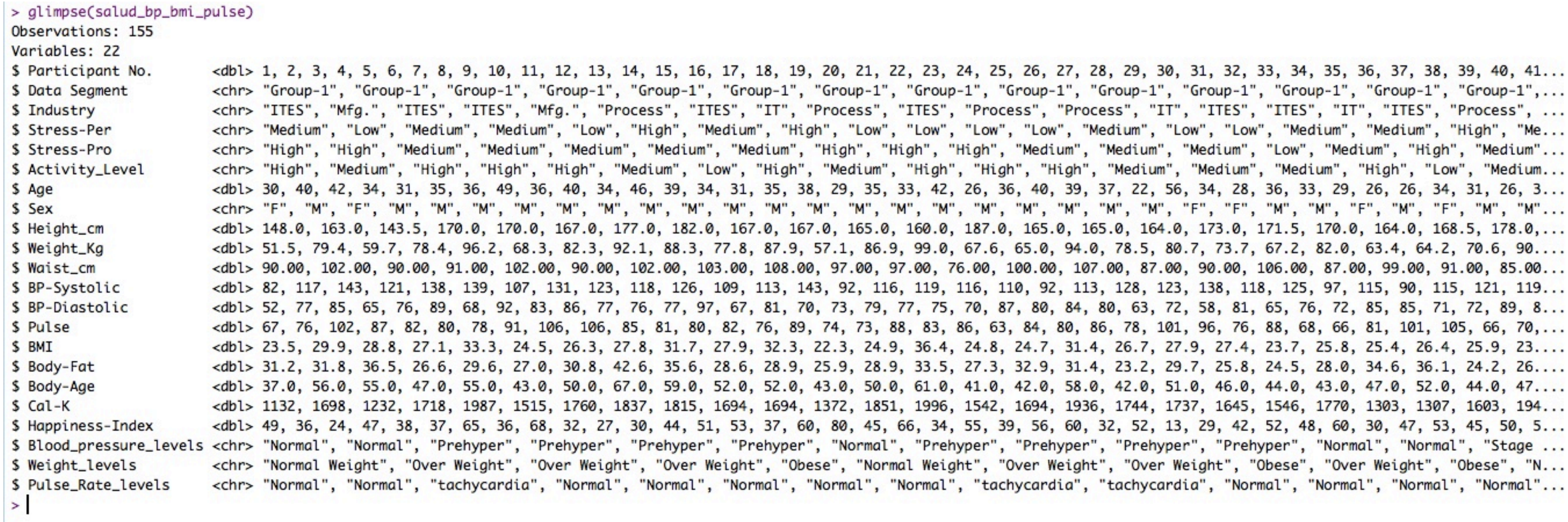
Exploratory Data Analysis of Health DataSet

Data Exploration Steps:

Pulse rate levels:

Pulse rate	Pulse_Rate_levels
Between(60,100)	Normal
>100	Tachycardia
<60	Bradycardia

Note: Here we can look at the data frame after getting transformed with the new variables



```
> glimpse(salud_bp_bmi_pulse)
Observations: 155
Variables: 22
$ Participant No.       <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41...
$ Data Segment          <chr> "Group-1", "Group-1", "Group-1", "Group-1", "Group-1", "Group-1", "Group-1", "Group-1", "Group-1", "Group-1", "Group-1", "Group-1", "Group-1", "Group-1",...
$ Industry              <chr> "ITES", "Mfg.", "ITES", "ITES", "Mfg.", "Process", "ITES", "IT", "Process", "ITES", "Process", "Process", "IT", "ITES", "ITES", "IT", "ITES", "Process", ...
$ Stress-Per            <chr> "Medium", "Low", "Medium", "Medium", "Low", "High", "Medium", "High", "Low", "Low", "Low", "Low", "Medium", "Low", "Low", "Medium", "Medium", "High", "Me...
$ Stress-Pro            <chr> "High", "High", "Medium", "Medium", "Medium", "Medium", "Medium", "High", "High", "High", "Medium", "Medium", "Medium", "Low", "Medium", "High", "Medium"...
$ Activity_Level        <chr> "High", "Medium", "High", "High", "High", "Medium", "Low", "High", "Medium", "High", "High", "High", "Medium", "Medium", "Medium", "High", "Low", "Medium...
$ Age                   <dbl> 30, 40, 42, 34, 31, 35, 36, 49, 36, 40, 34, 46, 39, 34, 31, 35, 38, 29, 35, 33, 42, 26, 36, 40, 39, 37, 22, 56, 34, 28, 36, 33, 29, 26, 26, 34, 31, 26, 3...
$ Sex                   <chr> "F", "M", "F", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "F", "F", "M", "M", "F", "M", "F", "M", "M"...
$ Height_cm             <dbl> 148.0, 163.0, 143.5, 170.0, 170.0, 167.0, 177.0, 182.0, 167.0, 167.0, 165.0, 160.0, 187.0, 165.0, 165.0, 164.0, 173.0, 171.5, 170.0, 164.0, 168.5, 178.0,...
$ Weight_Kg             <dbl> 51.5, 79.4, 59.7, 78.4, 96.2, 68.3, 82.3, 92.1, 88.3, 77.8, 87.9, 57.1, 86.9, 99.0, 67.6, 65.0, 94.0, 78.5, 80.7, 73.7, 67.2, 82.0, 63.4, 64.2, 70.6, 90....
$ Waist_cm              <dbl> 90.00, 102.00, 90.00, 91.00, 102.00, 90.00, 102.00, 103.00, 108.00, 97.00, 97.00, 76.00, 100.00, 107.00, 87.00, 90.00, 106.00, 87.00, 99.00, 91.00, 85.00...
$ BP-Systolic           <dbl> 82, 117, 143, 121, 138, 139, 107, 131, 123, 118, 126, 109, 113, 143, 92, 116, 119, 116, 110, 92, 113, 128, 123, 138, 118, 125, 97, 115, 90, 115, 121, 119...
$ BP-Diastolic          <dbl> 52, 77, 85, 65, 76, 89, 68, 92, 83, 86, 77, 76, 77, 97, 67, 81, 70, 73, 79, 77, 75, 70, 87, 80, 84, 80, 63, 72, 58, 81, 65, 76, 72, 85, 85, 71, 72, 89, 8...
$ Pulse                 <dbl> 67, 76, 102, 87, 82, 80, 78, 91, 106, 106, 85, 81, 80, 82, 76, 89, 74, 73, 88, 83, 86, 63, 84, 80, 86, 78, 101, 96, 76, 88, 68, 66, 81, 101, 105, 66, 70,...
$ BMI                   <dbl> 23.5, 29.9, 28.8, 27.1, 33.3, 24.5, 26.3, 27.8, 31.7, 27.9, 32.3, 22.3, 24.9, 36.4, 24.8, 24.7, 31.4, 26.7, 27.9, 27.4, 23.7, 25.8, 25.4, 26.4, 25.9, 23....
$ Body-Fat              <dbl> 31.2, 31.8, 36.5, 26.6, 29.6, 27.0, 30.8, 42.6, 35.6, 28.6, 28.9, 25.9, 28.9, 33.5, 27.3, 32.9, 31.4, 23.2, 29.7, 25.8, 24.5, 28.0, 34.6, 36.1, 24.2, 26....
$ Body-Age              <dbl> 37.0, 56.0, 55.0, 47.0, 55.0, 43.0, 50.0, 67.0, 59.0, 52.0, 52.0, 43.0, 50.0, 61.0, 41.0, 42.0, 58.0, 42.0, 51.0, 46.0, 44.0, 43.0, 47.0, 52.0, 44.0, 47....
$ Cal-K                 <dbl> 1132, 1698, 1232, 1718, 1987, 1515, 1760, 1837, 1815, 1694, 1694, 1372, 1851, 1996, 1542, 1694, 1936, 1744, 1737, 1645, 1546, 1770, 1303, 1307, 1603, 194...
$ Happiness-Index       <dbl> 49, 36, 24, 47, 38, 37, 65, 36, 68, 32, 27, 30, 44, 51, 53, 37, 60, 80, 45, 66, 34, 55, 39, 56, 60, 32, 52, 13, 29, 42, 52, 48, 60, 30, 47, 53, 45, 50, 5...
$ Blood_pressure_levels <chr> "Normal", "Normal", "Prehyper", "Prehyper", "Prehyper", "Prehyper", "Normal", "Prehyper", "Prehyper", "Prehyper", "Prehyper", "Normal", "Normal", "Stage ...
$ Weight_levels         <chr> "Normal Weight", "Over Weight", "Over Weight", "Over Weight", "Obese", "Normal Weight", "Over Weight", "Over Weight", "Obese", "Over Weight", "Obese", "N...
$ Pulse_Rate_levels     <chr> "Normal", "Normal", "tachycardia", "Normal", "Normal", "Normal", "Normal", "Normal", "tachycardia", "tachycardia", "Normal", "Normal", "Normal", "Normal"...
> |
```

Exploratory Data Analysis of Health DataSet

Data Exploration Steps:

Formulate questions:

1. How many enrolments have been done for the course Business analytics at ISI in each group?, is this course getting popularised as the days pass by? And are the female participants more than male participants ?

2. How are the head counts of different industries people ove the last 5 groups?, Which industry people are more interested in taking up the Business Analytics course at ISI to purse their career in Analytics? And are the academics people getting enrolled in this course ?.

3. What are the booming industries over the groups, at which type of industries the peoples are working more? And How are the stats of women workforce among these industries?

4. In which industries the professional stress levels are recorded at higher ?

5. How are the correlation between the different health measurements of the people?

6. What age of people have got higher BMI levels?. Is there any difference in BMI levels of Men and Woman?

7. How would be the BMI levels of people when they do physical activities daily ? Is there any change in BMI levels of people who do physical activities and people who don't ?

8. How are the stats of Blood pressure levels in Men and Women?, At what age the people are going to the stage of Pre-hyper ?

9. How are the weight levels in Men and Woman ?, At what age people are becoming more obese ?. Are there any woman under the Obese category ? And are there any underweight peoples in both of these genders?

10. How are the pulse rate levels in different ages of peoples ? , at what age the pulse rates are at high ? And are the tachycardia stage people more than the people of Normal stage ?

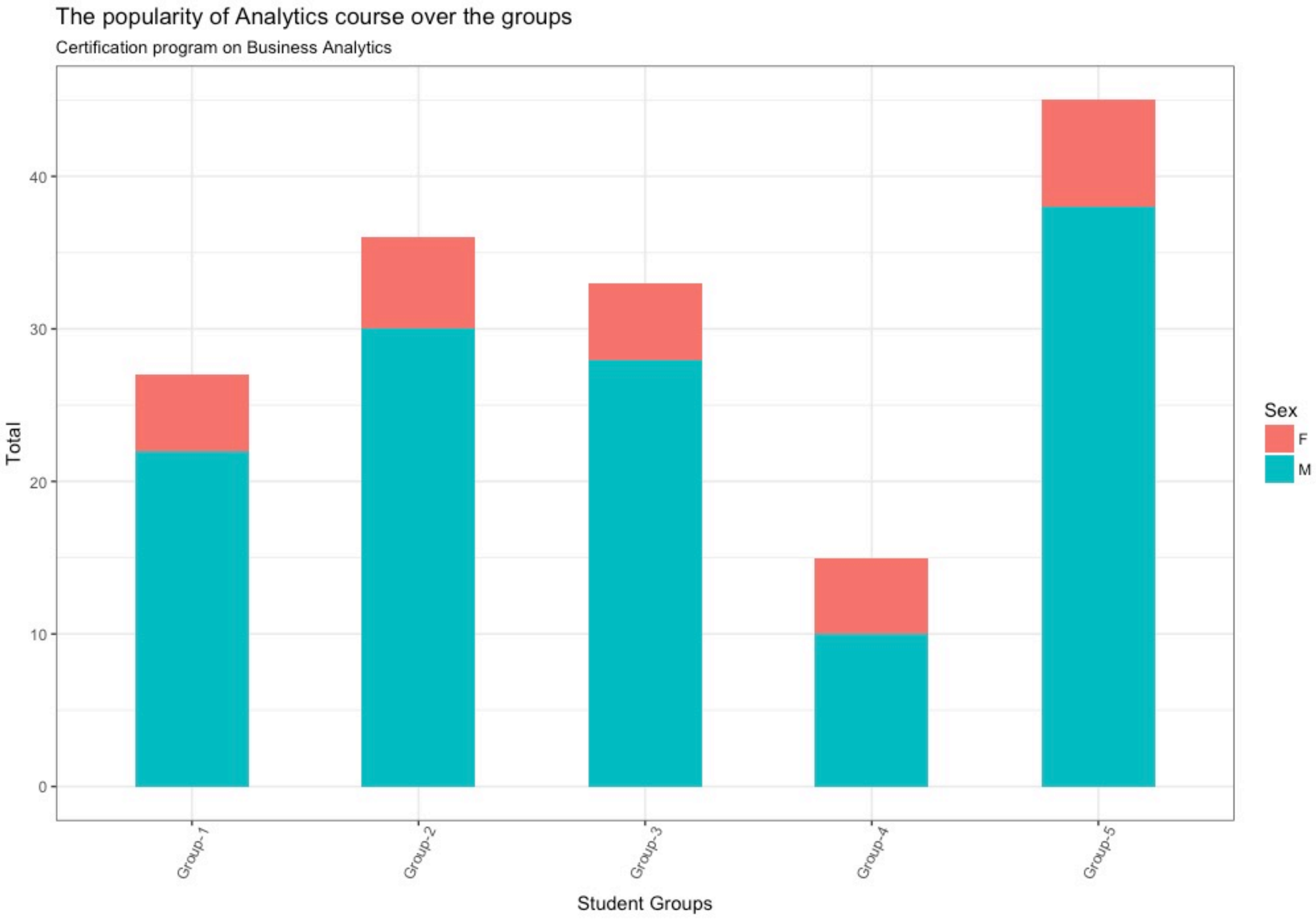
Exploratory Data Analysis of Health DataSet

Data Exploration Steps:

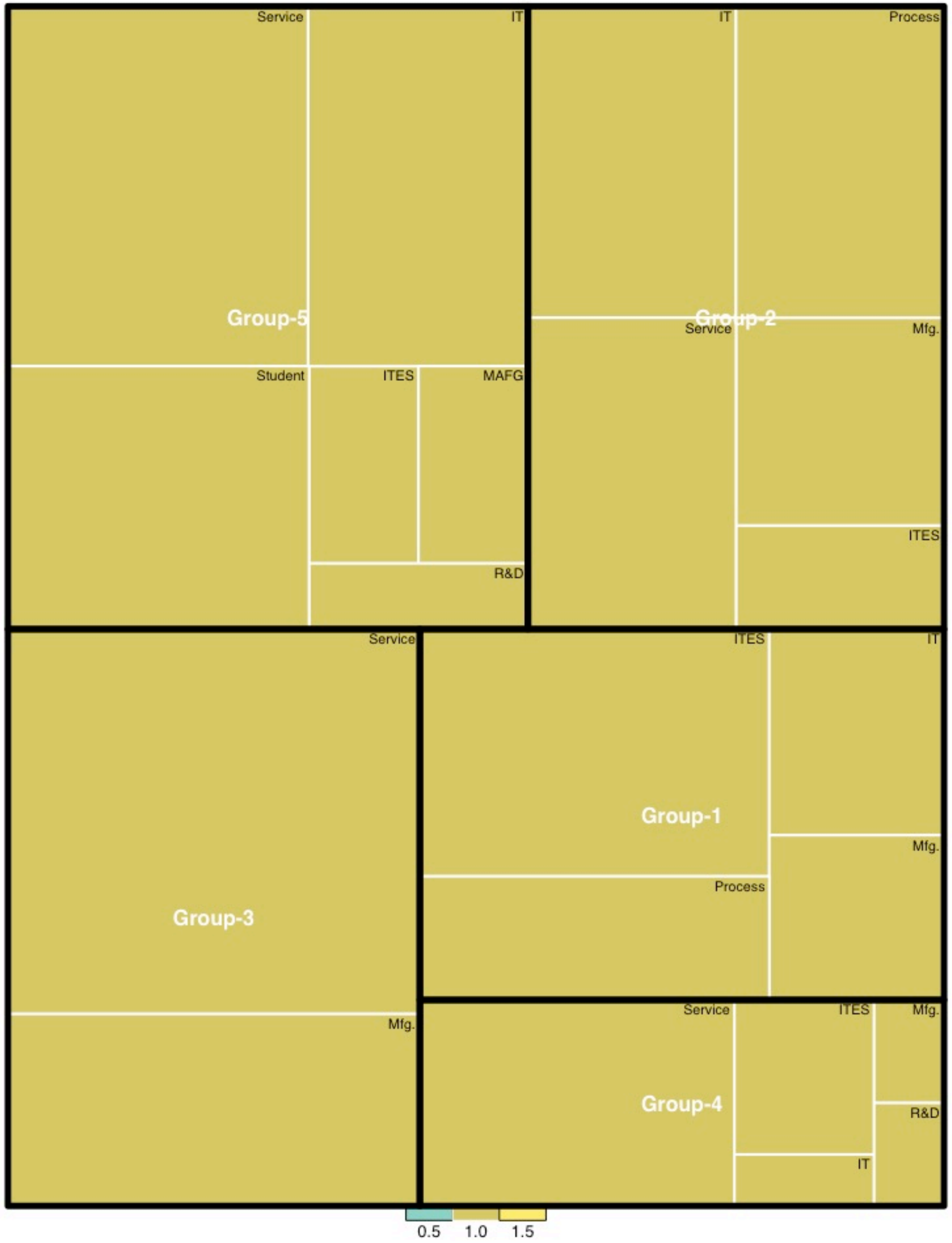
Data visualisation: Data visualisation is perhaps the fastest and most useful way to summarise and learn more about our data and here it helps us to answer the 10 formulated questions from the data set.

We can walk through each one of them in the next 10 slides.

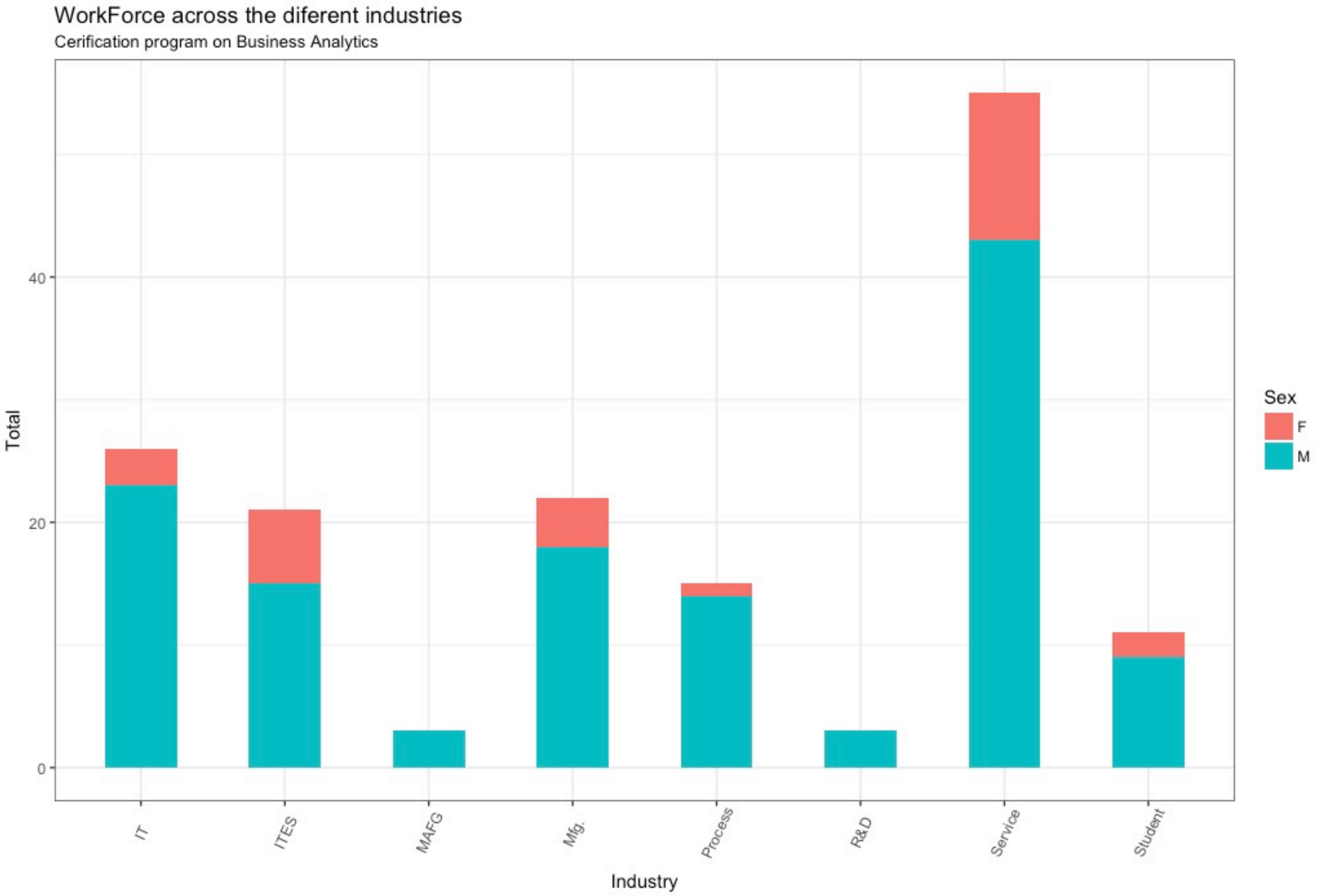
Answer 1: The popularity of the Business Analytics course at ISI-Hyderabad



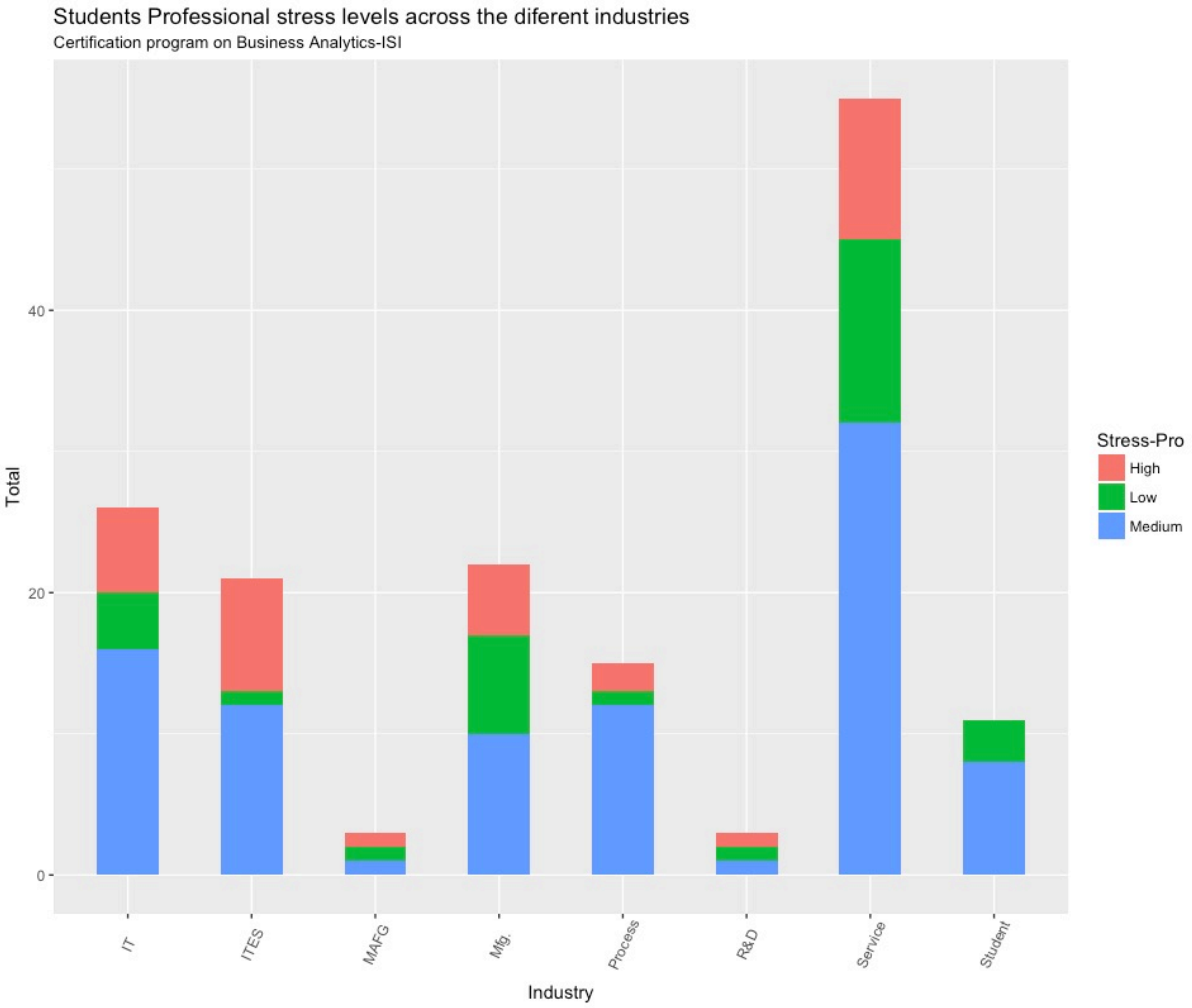
Answer 2: The different industries peoples are spread over the groups at ISI



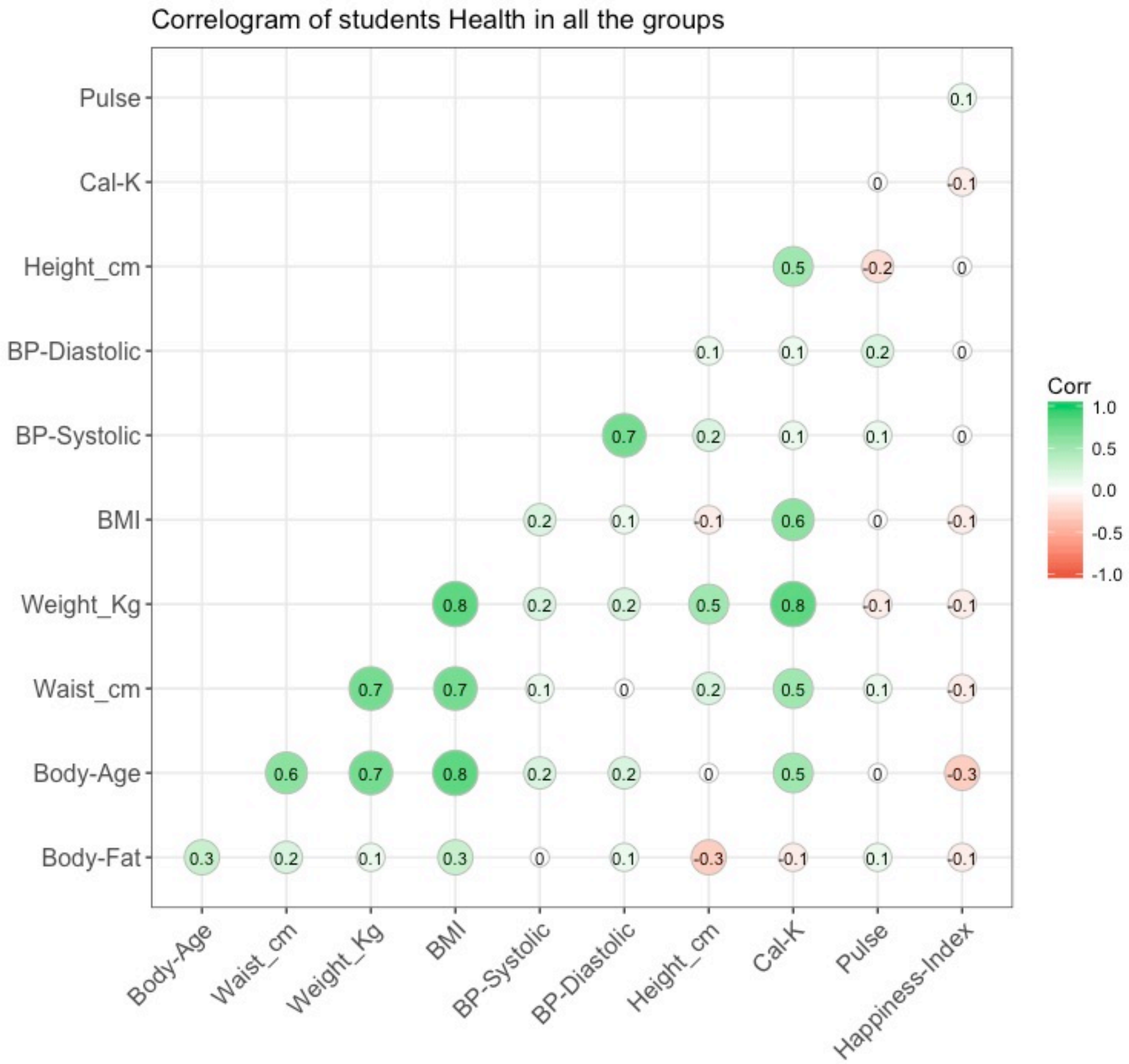
Answer 3: Workforce at the different industries



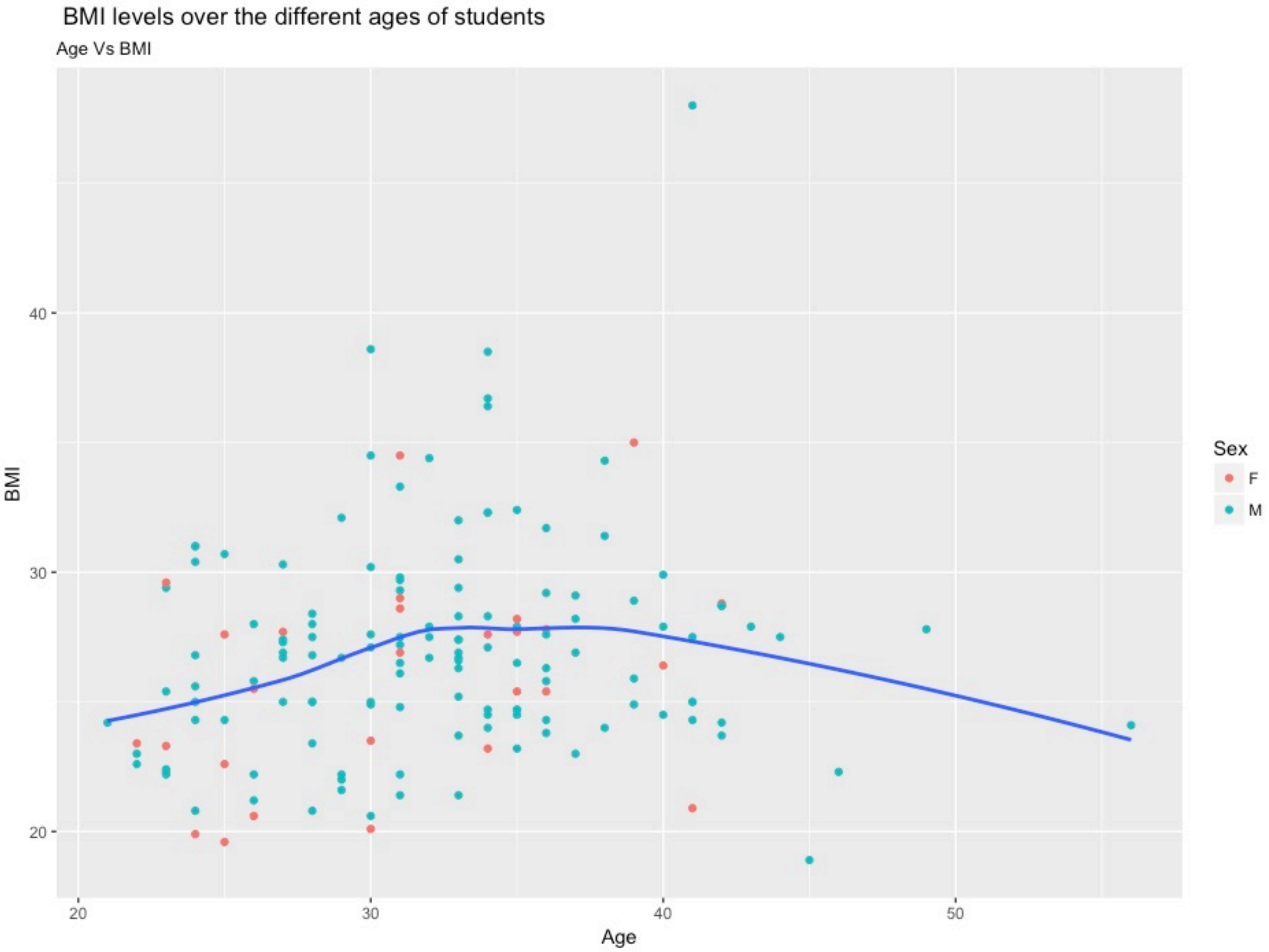
Answer 4: Professional stress levels across the industries



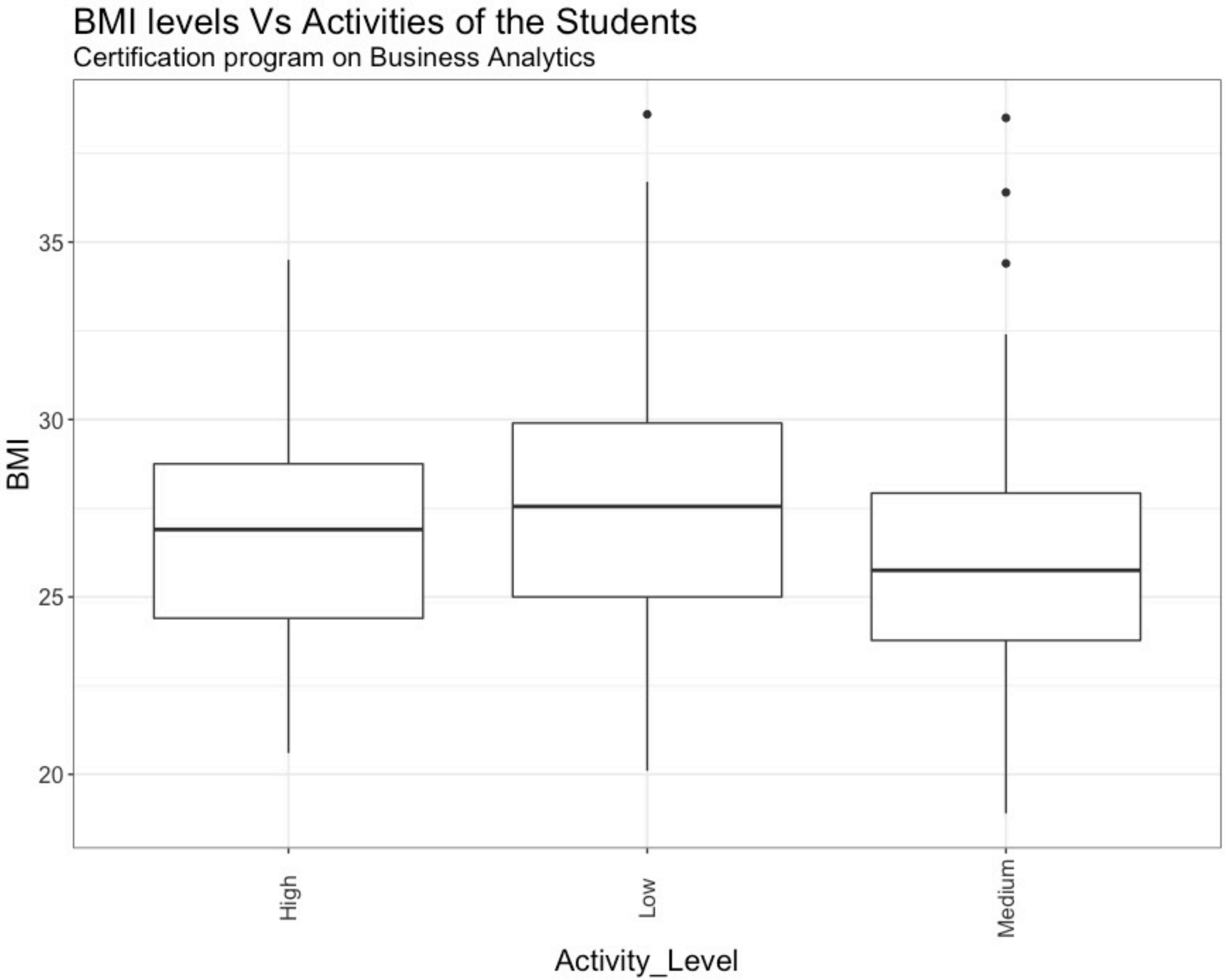
Answer 5: Correlation of peoples health stats



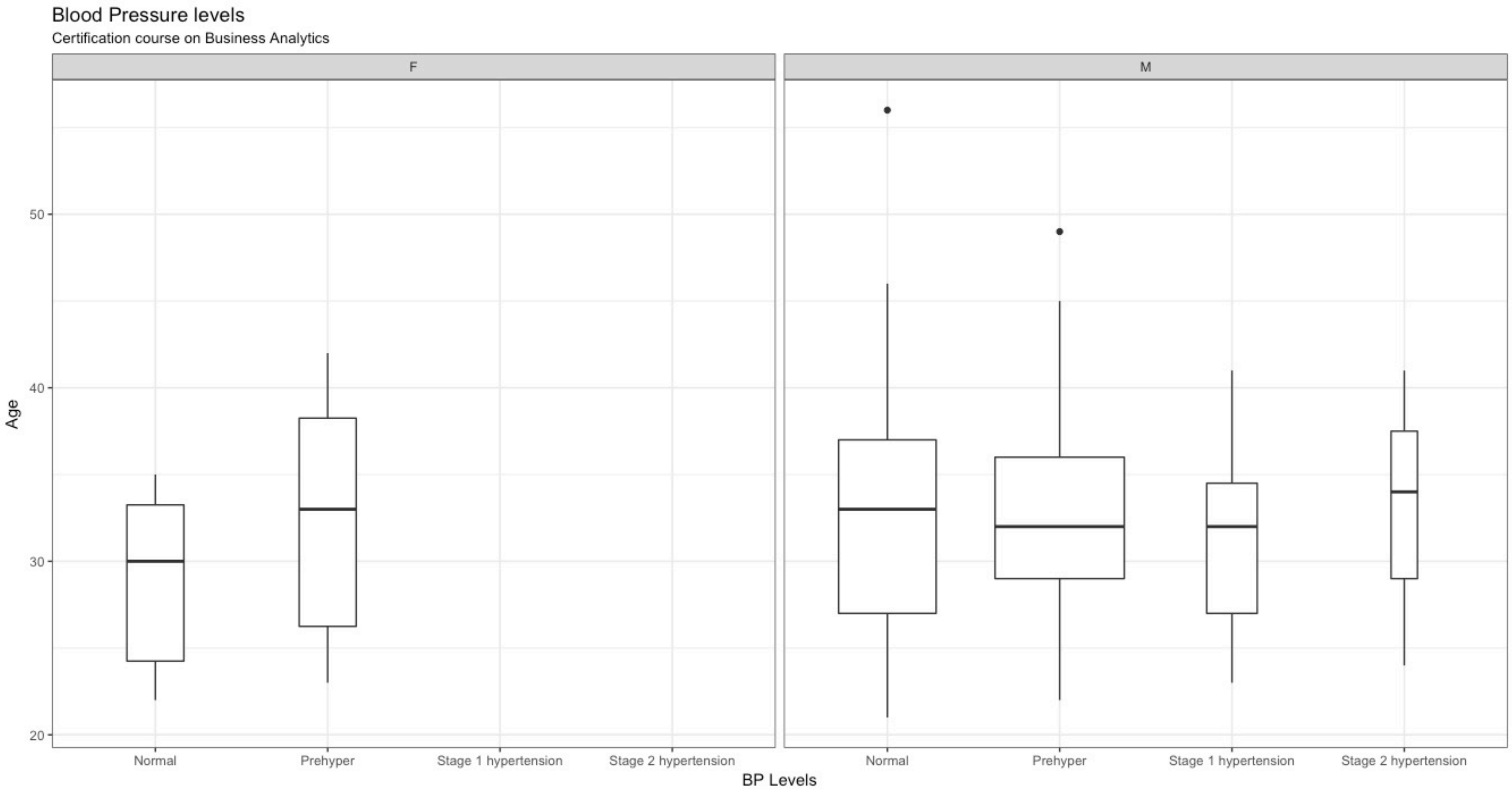
Answer 6: BMI levels over the different ages of students



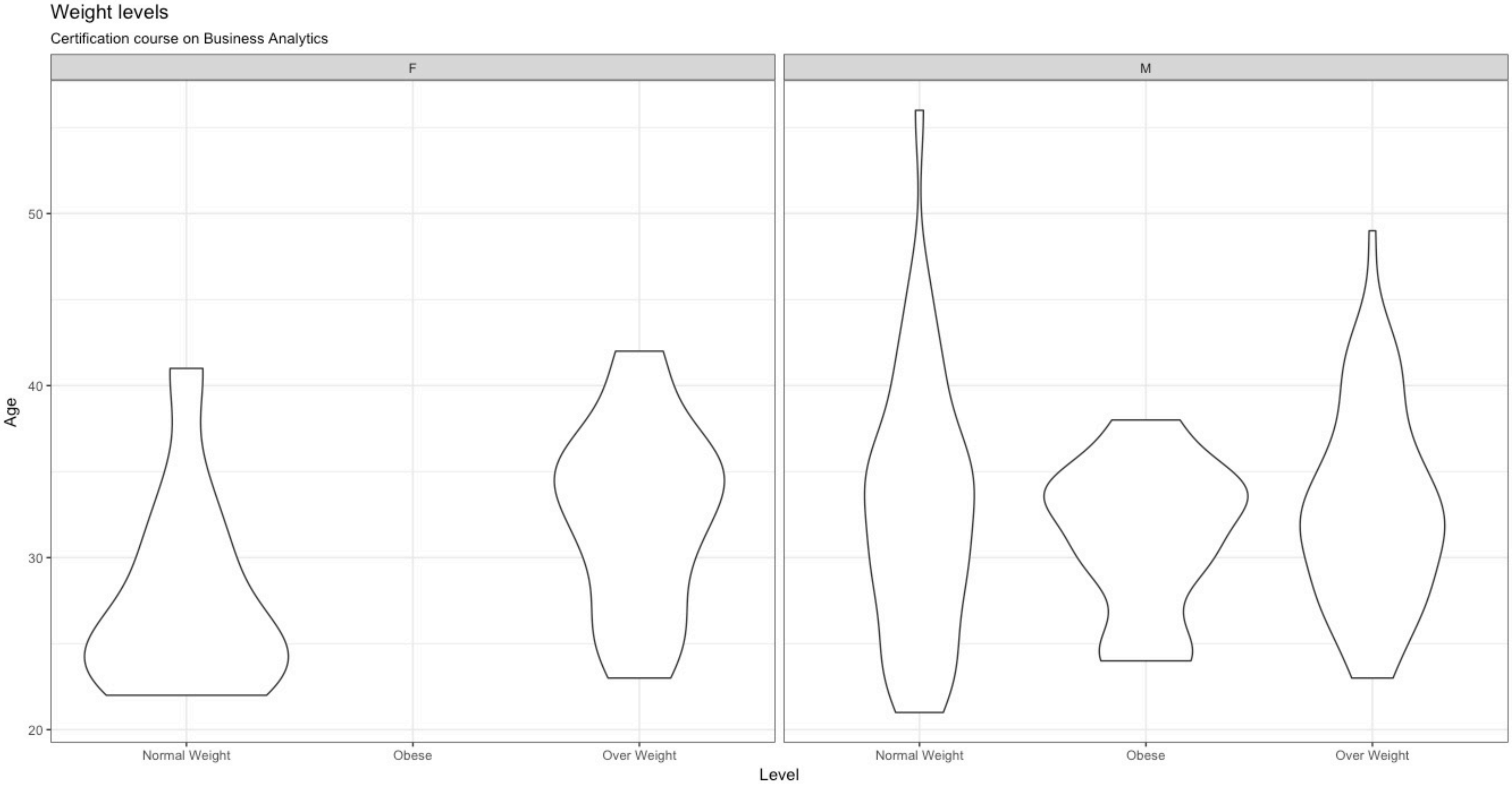
Answer 7: BMI levels among the people who do physical activities and people who don’t do



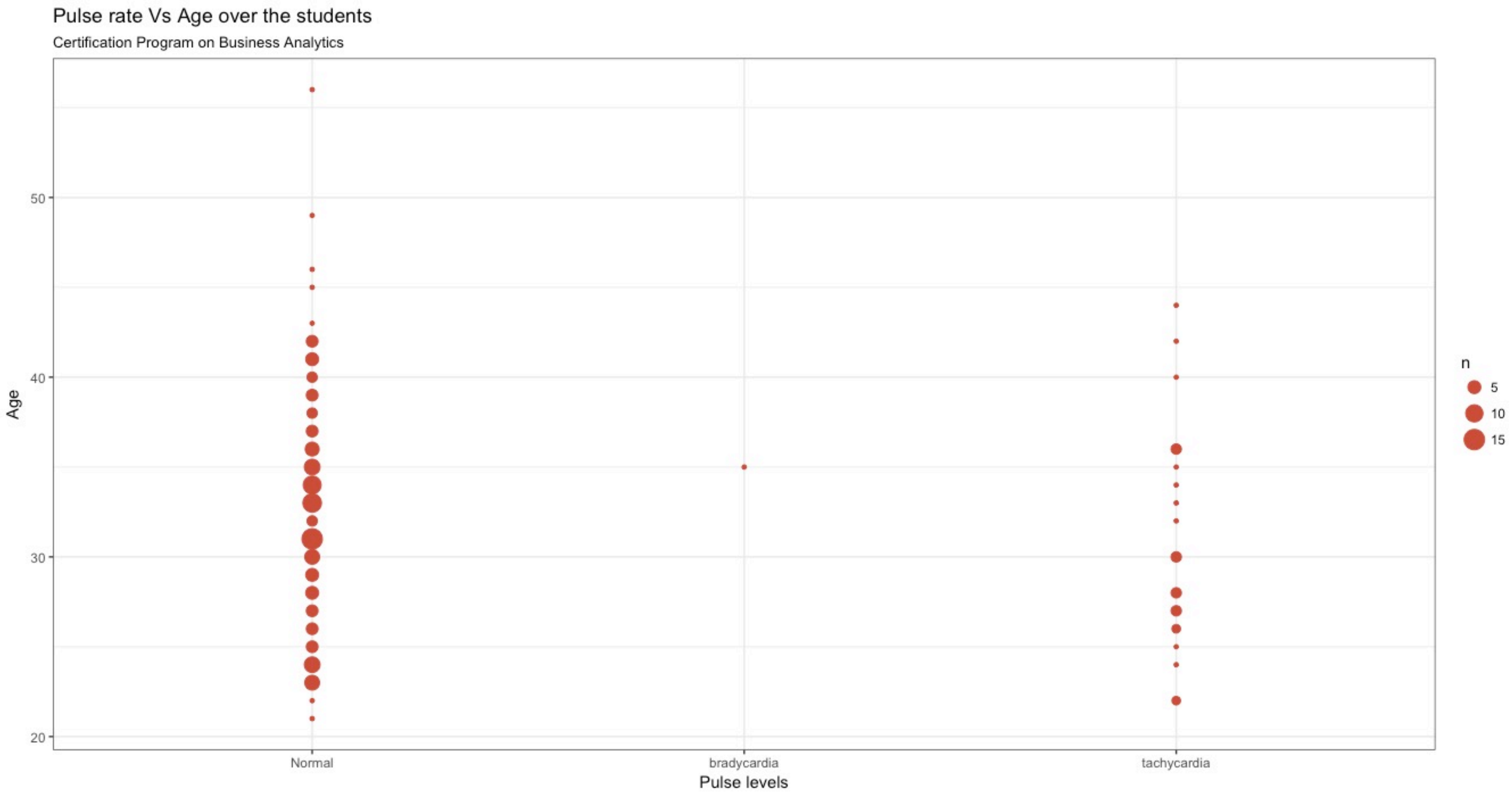
Answer 8: Categorisation of Blood Pressure levels



Answer 9: Categorisation of Weight levels



Answer 10: Categorisation of Pulse rate levels



Business Analytics problem

Trend analysis:

Trend analysis is the process of comparing data over time to identify any consistent results or trends. After looking through the health data set of about 155 observation I have come up with a business problem as written below,

Dataset has contained the information of candidates who have enrolled for a course in Group wise, like group 1, group 2..group5, here there is no information provided about when the each group has started, hence we are unable to find out the trends and wouldn’t predict whether the course would continue for the next years or not.

QFD &Kano Model

QFD:

Quality function deployment (QFD) is one of the very useful quality systems tools commonly applied to fulfil customer needs and improve customer satisfaction in many industries.

QFD is composed of four stages:

1. **Complete the house of quality (HOQ).**
2. **Design the product –** Determine tolerance of each part of the product so that it satisfies the target value identified from the HOQ.
3. **Design the process –** Determine the necessary production process that will satisfy tolerances established during product design.
4. **Control the process –** Determine quality standards for the new product design.

QFD offers a customer-oriented approach, supporting design teams in developing new products based on an assessment of customer needs. Customer needs are translated into design attributes, which are then deployed in process and quality requirements.

Benefits derived from using the QFD tool include:

- The creation of work teams that include multiple skills and experiences
- The determination of specific work aims
- A display of a wide variety of important design information in one place in a compact form
- Reduced overall costs from realising a reduction in design changes
- Reduced production costs by eliminating redundant features and over-design

QFD & Kano model

Completing the House of Quality

Step 1: Customer Requirements and Needs

Step 2: Ranking the Requirements

Step 3: Comparing Competitors

Step 4: Transforming Customer Requirements into Design Engineering Characteristics

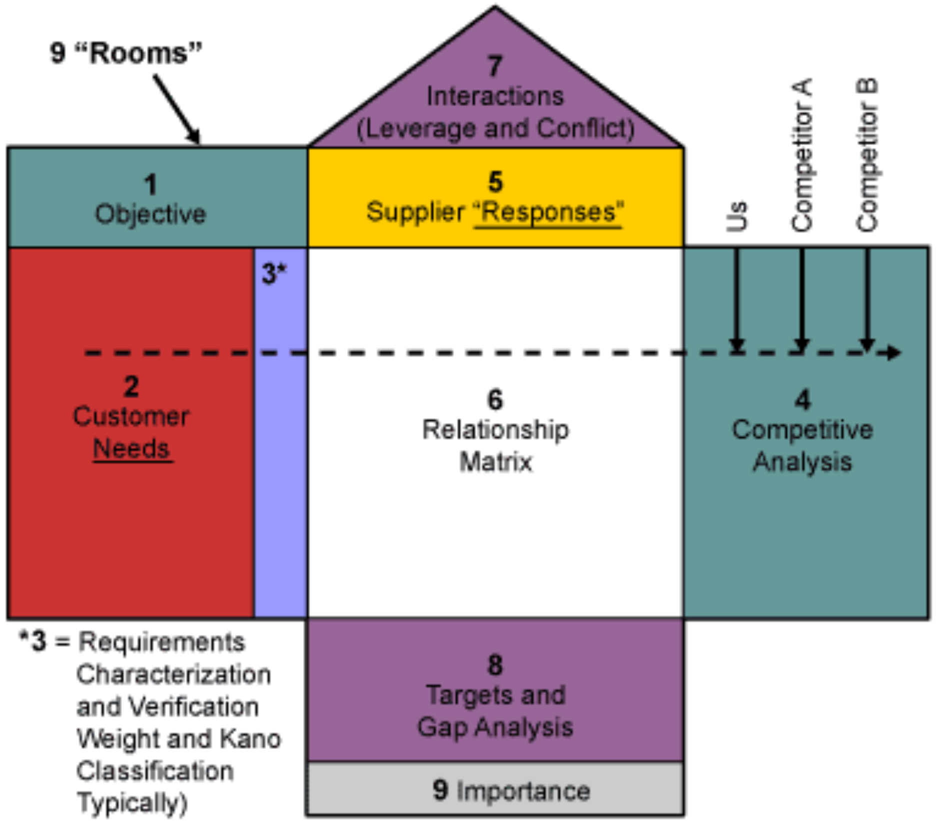
Step 5: Determining the Relationship Between Customer Requirements and Engineering Characteristics

Step 6: Comparing Competitor Characteristics with Design Engineering Characteristics

Step 7: Completing the HOQ Roof

Step 8: Weighing the Engineering Characteristics

Step 9: Determining New Target Values for Characteristics


9 "Rooms"
7 Interactions (Leverage and Conflict)
1 Objective
5 Supplier "Responses"
Us
Competitor A
Competitor B
3*
2 Customer Needs
6 Relationship Matrix
4 Competitive Analysis
*3 = Requirements Characterization and Verification Weight and Kano Classification Typically)
8 Targets and Gap Analysis
9 Importance

QFD and Kano model

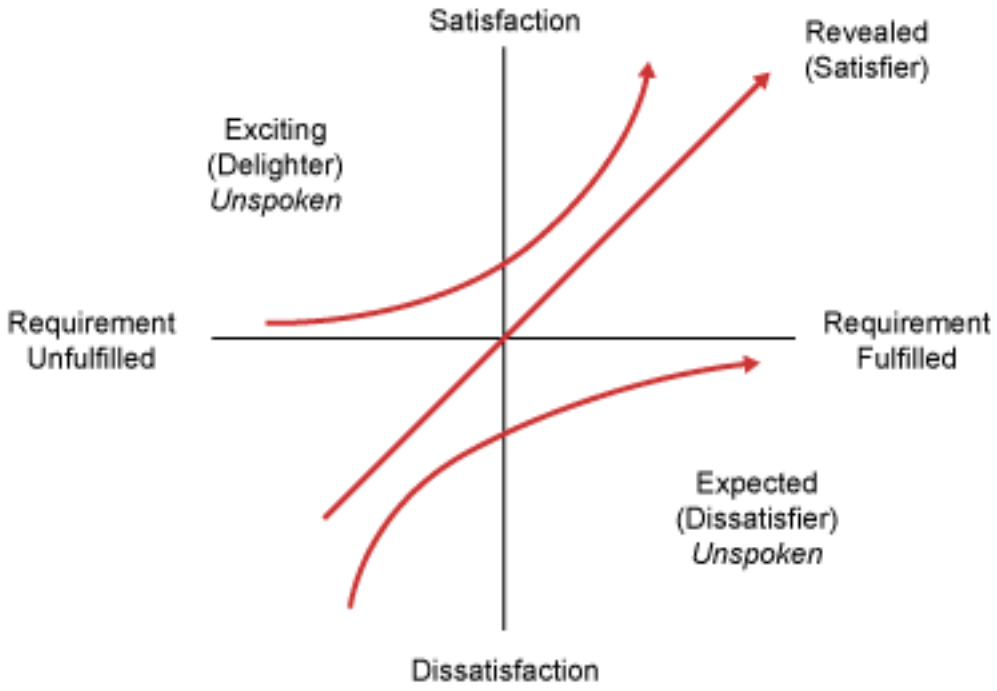
Kano Model of Customer Satisfaction:

The Kano model distinguishes between three types of product or service requirements that influence customer satisfaction,

1. Must-be requirements

3. One-dimensional requirements

5. Attractive requirements



QFD and Kano model

A quick glance at use case from my organisation:

We are offering a network protection service to one of telecom in south America, this product is specially designed for messaging security across their network and here they have observed that they continue to be at risk from signalling attacks including the tracking of VIP locations, interception of calls and texts, fraud, banking and 2FA cybercrime. Here there is an urgent need for them to avoid the dangers of attacks over SS7 and Diameter networks and in the industry no company provides a solution for controlling the attacks on SS7.

Then they have come to us with their handful of requirements to figure out a solution for their SS7 networks, their requirements have been looked through and validated by our various teams I.e Product Management, Software development etc. etc., finally we have taken a step forward to get a new product for SS7 that no one owns yet in the industry, we have spent around 2 years of time on getting it released to the client, and it has been names as SIGNALLING INTELLIGENCE (SIGIL) and deployed in their networks to tackle out the attacks on SS7.

Here we have fulfilled their requirement though it was very new for us and they were very glad to receive our services, since then we have become the world’s first global signalling intelligence and security analytics service company.

Thanks you (Muchas gracias)!!!!