National College of Ireland

Data Warehousing and Business Intelligence Project

on

Risk Factor of Heart Diseases in England and Wales

Bhumi Patel
18114865

MSc/PGDip Data Analytics – 2018/9

Submitted to: Dr. Horacio Gonzalez-Valez

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	Bhumi Patel
Student ID:	18114865
Programme:	MSc Data Analytics
Year:	2018/9
Module:	Data Warehousing and Business Intelligence
Lecturer:	Dr. Horacio Gonzalez-Valez
Submission Due Date:	26/11/2018
Project Title:	Risk Factor of Heart Diseases in England and Wales

I hereby certify that the information contained in this (my submission) is information pertaining to my own individual work that I conducted for this project. All information other than my own contribution is fully and appropriately referenced and listed in the relevant bibliography section. I assert that I have not referred to any work(s) other than those listed. I also include my TurnItIn report with this submission.

ALL materials used must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is an act of plagiarism and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	November 25, 2018

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Table 1: Mark sheet – do not edit

Criteria	Mark Awarded	Comment(s)
Objectives	of 5	
Related Work	of 10	
Data	of 25	
ETL	of 20	
Application	of 30	
Video	of 10	
Presentation	of 10	
Total	of 100	

Project Check List

This section capture the core requirements that the project entails represented as a check list for convenience.

- ☒ Used LaTeX template
- ☒ Three Business Requirements listed in introduction
- ☒ At least one structured data source
- ☒ At least one unstructured data source
- ☒ At least three sources of data
- ☒ Described all sources of data
- ☒ All sources of data are less than one year old, i.e. released after 17/09/2017
- ☒ Inserted and discussed star schema
- ☒ Completed logical data map
- ☒ Discussed the high level ETL strategy
- ☒ Provided 3 BI queries
- ☒ Detailed the sources of data used in each query
- ☒ Discussed the implications of results in each query
- ☒ Reviewed at least 5-10 appropriate papers on topic of your DWBI project

Risk Factor of Heart Diseases in England and Wales

Bhumi Patel
18114865

November 25, 2018

Abstract

Abstract goes here. You should provide a high-level (approx. 150 – 250 words) overview of your project, its motivation, and the core objectives / business requirements addressed.

1 Introduction

In this section you need to motivate your project (using citations). Why are you doing it? You should articulate the business requirements that your project seeks to address as motivation. These should be quantifiable and addressed via the BI Queries in Section 7.

Lorem ipsum dolor sit amet, ut veri deleniti eloquentiam sea Feng & Buyya (2016). Ea commodo aperiam complectitur pri, usu et case dolore. Kune et al. (2016) ad quidam regione percipitur, est ut possit bonorum persecuti. Quis utinam offendit eu usu, eu accumsan disputando per, id cibo reprehendunt sit Beloglazov & Buyya (2015), Gomes et al. (2015). In melius legendos corrumpit pro. Eos dico dignissim voluptatibus et, duo nisl cibo ut. Diceret periculis posidonium cum eu. Gomes et al. (2015) regione nam ex. Vix id viris phaedrum. Pri augue cetero probatus ut.

(Req-1) my first requirement

(Req-2) my second requirement

(Req-3) ...

2 Data Sources

There are 4 wellspring of datasets which is exploited in the project.

2.1 Source 1 : Statista

2 structured datasets were taken for implementing the project which shows the risk factors of Heart diseases.

- **Statista : Overweight prevalence**
  Release date:December 2017
  This dataset is downloaded in xlsx format from : `https://www.statista.com/`

Source	Type	Brief Summary
Statista	Structured	The platform combines economic data, consumer insights,sentiment surveys,reviews and markets research trends.which are genuine hence I have used Obesity rate, Overweight rate and winter death from three different sources.
Office for National Statistic	Structured	O.N.S is a largest independent provider of U.K and gives government statistics of the U.K which helped me to capture data on Deaths from diseases of the cardiovascular system and ischaemic heart disease,
Wikipedia	Structured	Wikipedia provides the information about demography of Wales and England.Its gives the brief idea about population and death based on year which is further helpful in my business query.
Guardian	Unstructured	The Guardian is a highest circulatory newspaper in U.K which proposed me to utilize opinion and reviews of people on fast food and junk food.

Table 2: Summary of sources of data used in the project

`statistics/334085/overweight-prevalence-england/`. The above URL contains 2 worksheets, includes description of data and the overweight prevalence among grown-upin England from 2000 to 2016.There are 3 columns and 17 rows which shows overweight prevalence by gender in percentage.

Information about overweight rate and other relevant columns are included in the business query which contains percentage of overweight based on gender and Year in England.

- **Statista : Obese prevalence**
  Release date:December 2017
  This dataset is downloaded in xlsx format From : `https://www.statista.com/statistics/376074/obesity-prevalence-by-gender-in-england-uk/`. The above URL contains 2 worksheets, includes description of data and the prevalence of adults who are obese in England from 2003 to 2016.There are 3 columns and 14 rows which shows obesity prevalence by gender in percentage.

  Information about obesity rate and other relevant columns are included in the business query which contains percentage of obesity based on gender and Year in England.

- **Statista : Winter Deaths by age group and gender**
  Release date:November 2017
  This dataset is downloaded in xlsx format From : `https://www.statista.com/statistics/283018/excess-winter-deaths-in-england-wales-by-age-and-gender/`. The above URL contains 2 worksheets, which includes description of data and the

number of deaths in winter 2016/2017 based on age group and gender for England and wales .There are 3 columns and 4 rows which shows winter deaths in numbers.

Information about winter deaths and other relevant columns are included in the business query which contains numbers of deaths based on gender and age in England and Wales.

2.2 Source 2 : Office for National Statistic

Release date: 7 December 2017
The structured dataset is downloaded in xlsx format from ONS site: `https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/adhocs/007824deathsfromdiseasesofthecardiovascularsystemandischaemicheartdisease2010to2016` .The above URL contains 5 worksheets,which includes Notes/Terms and Conditions with contact details, Table of content, diseases name with Tenth Revision code (ICD-10) and it also contains 75 columns of information on 2 diseases and 20 rows which show the information in number.This is the structured data containing deaths from the diseases of cardiovascular system and ischaemic heart diseases in England, Wales including non-residents.

Information about diseases and other relevant columns are included in the business query which contains number of death ratio based on gender,age group and month in England Wales.

2.3 Source 3 : Wikipedia

2 semi structured datasets are used for my project and are extracted by using R programming.

- Wikipedia : Demography of England
  This dataset is extracted from : `https://en.wikipedia.org/wiki/Demography_of_England` The above URL contains Demography of England from 1940 to 2017.There are 3 columns and 77 rows which shows average population and Total Deaths in number.

  Information of the Avg population and Deaths rate are included in the business query which contains number based on year in England.

- Wikipedia : Demography of Wales
  This dataset is extracted from :
  `https://en.wikipedia.org/wiki/Demography_of_Wales`. The above URL contains Demography of Wales from 1940 to 2017. There are 3 columns and 77 rows which shows average population and Total Deaths in number.

  All Information of the are included in the business query which contains number based year in Wales.

2.4 Source 4 : Guardian

This unstructured datasets is extracted from : `https://www.theguardian.com` For Unstructured source I have extracted guardian data to perform sentiment analysis using API.

The sentiment count from guardian is used for visualizing preferences of junk food/fast food.

3 Related Work

In this section, discuss related work on your topic of choice. Consider answering the following questions:

Q1 How have these (or similar) datasets been used before?

Q2 What is generally known about the domain within which your requirements (in Section 1) are situated?

Q3 What significant results exist in this area, and how to you expect to add to them by undertaking this project?

See Hall et al. (2018) for an example of a lit review that looks at specific challenges and approaches within a given area.

4 Data Model

Data model is logical inter-relationships between data elements.In this project there are five dimension tables which is directly connected with fact table.The detail description of fact table and dimension table are represented in form of star schema,as star schema is more flexible and does not require normalized data.Attributes of star schema is shown below.

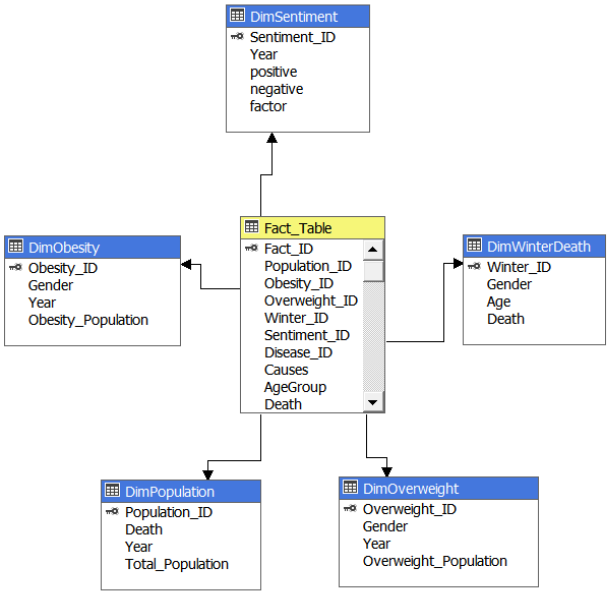

Figure 1: star schema

- **DimObesity**
  The dimObesity consist of multiple columns which provide information regarding the total number of people suffering from obesity by gender per year. Total number of people suffering from obesity is used for comparing with overweight count to see its total influence on population which was taken from dimoverwight and dim population respectively.Furthermore,it is connected with DimSentiment which gives sentiment analysis.Year and Gender is correlated with Fact table to form business query.Year and Gender data is used in this dimension to avoid redundancy.The attributes and values of dimobesity are taken from statista.

- **DimOverweight**
  The dimOverweight provide information about total number of people suffering from overweight by gender per year i.e. 2010 to 2016.This dimension is linked with other dimensions and fact table for forming BI query.Year and Gender data is used in this dimension to remove redundancy.The columns and values of dimOverweight are downloaded from statista.

- **DimPopulation**
  DimPopulation hold the total population and deaths of England and Wales.This dimension is interlinked with DimObesity, DimOverweight and Fact table by year,this helps me to find people suffering from obesity and overweight.Furthermore,it also gives number of deaths for heart diseases, measure for which is obtained from fact table.Year data value is used in this dimension to avoid duplicacy.The columns and values of dimPopulation are taken from Wikipedia.

- **DimWinterDeath**
  DimWinterDeath consists the attributes of Age, Gender and total number of deaths in winter.This dimension is linked with fact table by gender and age group to reduce redundancy and provide the information of death related to diseases and total winter death based on gender and age group.Values for this dimension is fetched from statista.

- **DimSentiment**
  Data for Dimsentiment is obtained by Guardian API for sentiment analysis.This dimension includes factor, positive and negative count based on year which is connected with the fact table by removing duplicate value. Dimsentiment is interlinked with other dimensions and fact table to give accurate result based on sentiment to see the total number people influenced by junk food and fast food which leads to obesity or overweight.

- **Fact_Table**
  Fact Table consist of both measure and dimension value having monthly grain.These measures and dimension are used to analyzed BI query.Fact table consist of Causes, age group, month, total deaths based on year which is interlink with other dimension table.

5 Logical Data Map

"Logical data model is used to describe the data in details without physically implemented in database through ETL process.Data mapping, maps two unique data models by creating data element mappings.In my project I have used three type of datasets i.e structure, semi-structure, unstructured and performing ETL process to reduce redundancies.The detail description of each attributes are shown below.

Table 3: Logical Data Map describing all transformations, sources and destinations for all components of the data model illustrated in Figure 2

Source	**Column**	**Destination**	**Column**	**Type**	**Transformation**
1	Men	DimObesity	Gender	Dimension	Transformed as character column value by gathering gender male
1	Women	DimObesity	Gender	Dimension	Transformed as character column value by gathering gender female
1	Men,Women	DimObesity	Gender	Dimension	Transformed as single character column value and by gathering gender in different column
1	Values	DimObesity	Obesity_ Population	Dimension	Using mathematical Transformed as single numeric column value and by gathering number of Overweight population in different column
1	Year	DimObesity	Year	Dimension	removed unwanted data and used 2010 to 2016
1	Women	DimOverweight	Gender	Dimension	Transformed as character column value by gathering gender female
1	Men	DimOverweight	Gender	Dimension	Transformed as character column value by gathering gender male
1	Year	DimOverweight	Year	Dimension	removed unwanted data and used 2010 to 2016
1	Women,Men	DimOverweight	Gender	Dimension	Transformed as single character column value and by gathering gender in different column

Continued on next page

Table 3 – *Continued from previous page*

Source	Column	Destination	Column	Type	Transformation
1	No columns name	DimOverweight	Overweight Population	Dimension	Using mathematical Transformed as single numeric column value and by gathering number of Overweight population in different column
1	Female	DimWinter Death	Gender	Dimension	Transformed as character column value by gathering gender female
1	Male	DimWinter Death	Gender	Dimension	Transformed as character column value by gathering gender male
1	Female,Male	DimWinter Death	Gender	Dimension	Transformed as single character column value and by gathering gender in different column
1	No Columns Name	DimWinter Death	Age	Dimension	Transformed as numeric value by gathering Age group based on gender
1	Death	DimWinter Death	Death	Dimension	Transformed as single numeric column value and by gathering number of deaths in different column
2	Sex	FactTable	Gender	Fact	converted into character by gathering values
2	Age Group	FactTable	AgeGroup	Fact	Remove total Age and convert into numeric formate by gathering values
2	20102016	Fact Table	Year	Fact	Transformed as single numeric column value and by gathering month wise in different column
2	JanFeb	FactTable	Month	Fact	Transformed as single numeric column value by gathering monthly data on the basis of death rate in different column
2	Causes	FactTable	Causes	Fact	Gathered two country of data and represent type of Diseases
2	Deaths	FactTable	Death	Dimension	summed up death rate of two columns into single column
3	England Population	DimPopulation	Total_ Population	Dimension	remove Null value and add into Total Population

Continued on next page

Table 3 – *Continued from previous page*

Source	**Column**	**Destination**	**Column**	**Type**	**Transformation**
3	England Deaths	DimPopulation	Deaths	Dimension	Its represent total deaths based on Year in England and added into Deaths
3	Wales population	DimPopulation	Total_ Population	Dimension	remove null value and add into Total Population
3	Wales Deaths	DimPopulation	Deaths	Dimension	Its represent total deaths based on Year in Wales and added into Deaths
3	England Population,Wales population	DimPopulation	Total_ Population	Dimension	.Its represent total population of England and Wales
3	England Deaths,Wales Deaths	DimPopulation	Death	Dimension	Converted into numeric form by using mathematical formula which represent total deaths in England and Wales
4	Factor	DimSentiment	factor	Dimension	Used only Unhealthy data remove unwanted data
4	Positive	DimSentiment	positive	Dimension	Merged positive sentiment data based on Year
4	Negative	DimSentiment	negative	Dimension	Merged negative sentiment data based on Year

6 ETL Process

The process of extracting information from source framework and carrying it into the data warehouse is commonly called as ETL, which stands for extraction, transformation and loading.Where important step is to extract or collect data from data source after that transform it into a staging area and at the end loaded into target file. One click automation is being implemented in which the data is fetched from flat data files and then cleaned, transformed and loaded into the staging area. Further to add to this process, dimensions and fact table are also being populated. All the tables are dropped every time the SSIS is executed and new tables are getting generated. This is done so that redundancy of data can be avoided and result can be made more accurate. Cube deployment is also done using this automation.

- **Extraction**
  Data extraction is a process of recovering data from structure, semi-structure and unstructured data source for further data processing.

  In this project various data source are used to perform business query. All the data were downloaded in xlsx format. This data was containing unwanted rows, columns and junk of information which was removed and only the necessary information was being extracted. In this case, a CSV file was written at the end of the cleaning process which was then loaded in the staging area.

  Similarly for semi-structure data, Wikipedia has been used to fetch the data using R programming. Wikipedia provided information in HTML content, which was supposed to be extracted. Bits and pieces of information about the same was extracted using a R library dplyr, ... Once the data was extracted, a final CSV file is then formed which is then loaded in the staging area.

  Likewise, for setting the needs of unstructured data The Guardian news API is used. The Guardian being one of the renowned news broadcaster in the United Kingdom, have provided an open source API for developers. This REST API provides the content is JSON format and the content is fetched using R code. Libraries like ..... are involved in this process. Sentiments analysis is then carried out to calculate the emotions present in the content. These sentiments are then used for the purpose of analysis.

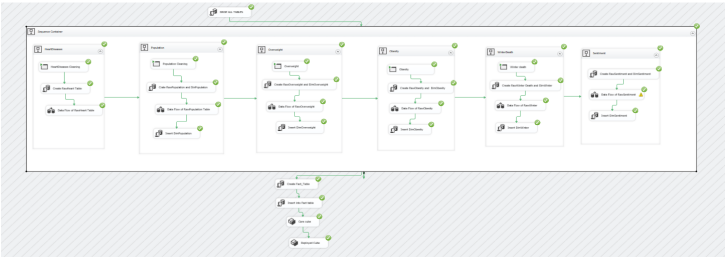

Figure 2: star schema

- **Transformation**
  The next step of the ETL cycle is the Transformation. In this process data is syncronized as per the need to the application. The content which is being extracted

from the data sources is then transform to provide the necessary informtaion over which SQL can be implemented. The first step towards the data transformation is to remove null values and special character from all the data sets. Merge, gather and spread function of R are being used on the columns and rows to align the data as per the need. Arithmetic operations are being performed over the numeric data to archive accurate result for BI query.

In the case of unstructured data, content provided by The Guardian API is in textual format. This content is fetched and NRC sentiment analysis is carried over it. Numeric values of sentiments are retrieved which can be used for further processing. Data is then written in a CSV file which is then used for the loading operation.

- **Loading**
  Loading is the last step of the ETL process in which chunk of data which is formulated after the cleaning and transformation is loaded in the data warehouse. Staging area is initially populated and further the data is transferred to the Data warehouse. There are 5 dimensions which are initially populated and then the fact table is populated with the primary keys of every dimension and its measures. Dimensions of Obesity, Overweight, Population, WinterDeath and Sentiment are getting formed. Join operation is performed and then data is being populated in the fact table.

  When the fact table is populated, formation of cube takes place. This cube is then deployed and can be used for visualization purposes.

7 Application

After deploying cube in SSAS, tableau software is used to analyze business intelligence query for data visualization.

7.1 BI Query 1:How Does unhealthy food correlate with risk factors of heart diseases?

This Query make use of 2 data sources(structure and unstructured) which comes from Statista and Guardian. Statista consists two different datasets based on prevalence of Obesity and Overweight,whereas Guardian shows positive and negative count of unhealthy diet and eating habits such as junk food and fast food preferred by people per year.To illustrate the above, four different attributes(positive ,negative, obese people and overweight people,year) are used from 3 data sets based on year.

From the lines graph representation, it clearly analyzed that people prefer more junk food and fast food per year and due to which there is an increase in overweight and obesity ratio among people shown in the other graph. So, it is analyzed that more number of people who are unhealthy mostly prefer junk food leading to increase in obesity and overweight problems which ultimately leads to increase in risk factor of heart diseases among the unhealthy people.

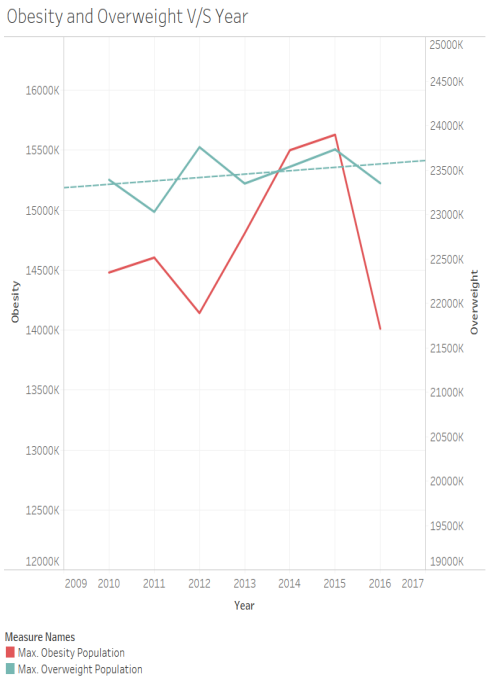

Figure 3: Risk factors per year

Figure 4: Sentiment analysis

7.2 BI Query 2: Which gender of population is suffering from Obesity and Overweight from 2010 to 2016?

For this query,the contributing source of data are structure and semi structured, which comes from Statista and Wikipedia. Statista consists two different datasets based on prevalence of Obesity and Overweight,whereas Wikipedia shows total population per year.To illustrate the above, four different attributes(population ,Gender, obese people and overweight people,year) are used from 4 data sets based on year.

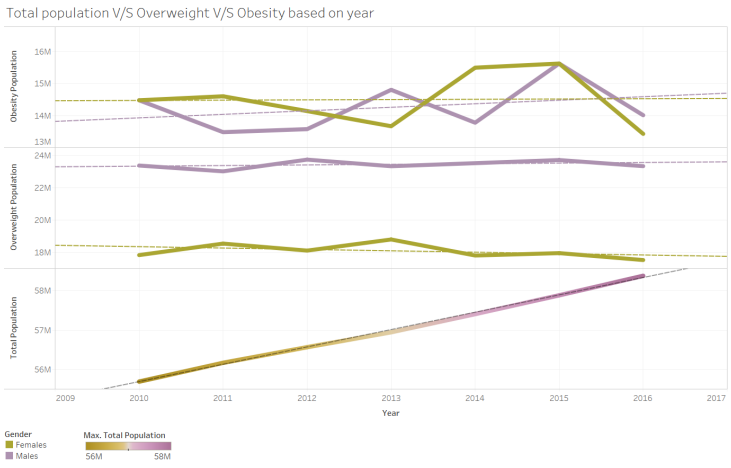

Figure 5: population with risk factor

Line graph is used to analyzed that from total population,males have higher number of overweight rate when compare with females in each year,whereas in obese the prevalence of female is slightly higher compared to male per year.To conclude,that more number

of females are suffering through obese and more number of males are suffering through overweight over the total population in England and Wales from the year 2010 to 2016.

7.3 BI Query 3:How many number of deaths is caused due to heart diseases?

This Query touches 2 data sources(structure and semi structured) which comes from Office for National Statistic(O.N.S) and Wikipedia.O.N.S consists total number of death caused due to heart diseases per year,whereas Wikipedia shows total number of deaths per year.To illustrate the above,three different attributes(Deaths,Diseases death, Year) are used from 3 data sets among which O.N.S consist of one data sets and Wikipedia consist of two data sets based on year in England and Wales.

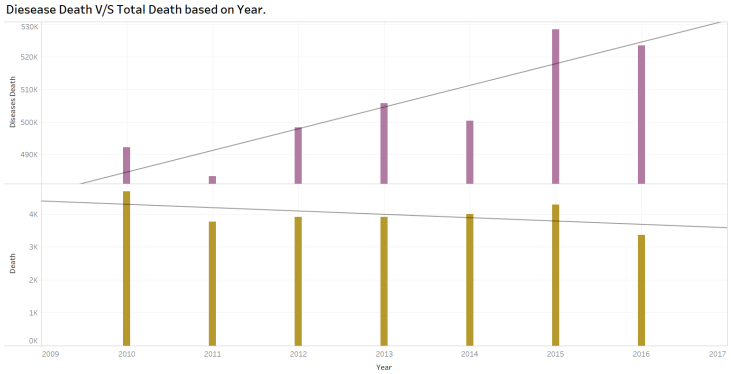

Figure 6: Diseases death ratio

A Bar graph representation is shown which helps to clearly analyze that more than 4000 peoples died due to heart diseases over total death in England and Wales from 2010 to 2016.

7.4 BI Query 4:Which age group and gender has more death ratio for heart related diseases in winter 2016?

This Query is fulfulled by using 2 data sources which comes from Office for National Statistic and Statista (O.N.S). Dataset from O.N.S consists of total number of deaths based on heart diseases by gender and age group per months,whereas Statista shows total number of winter deaths based on gender.To illustrate the above,five different attributes(Deaths,Diseases death, age group,gender, Causes) are used.

For the above figure representation,size of circles depends on number of deaths related to diseases when compared from total number of deaths in England and Wales for the year 2016 by gender and age group.It can be clearly analyzed that females of the age group 85 and above has larger number death ratio, followed by males of the same age group due to circulatory heart diseases.

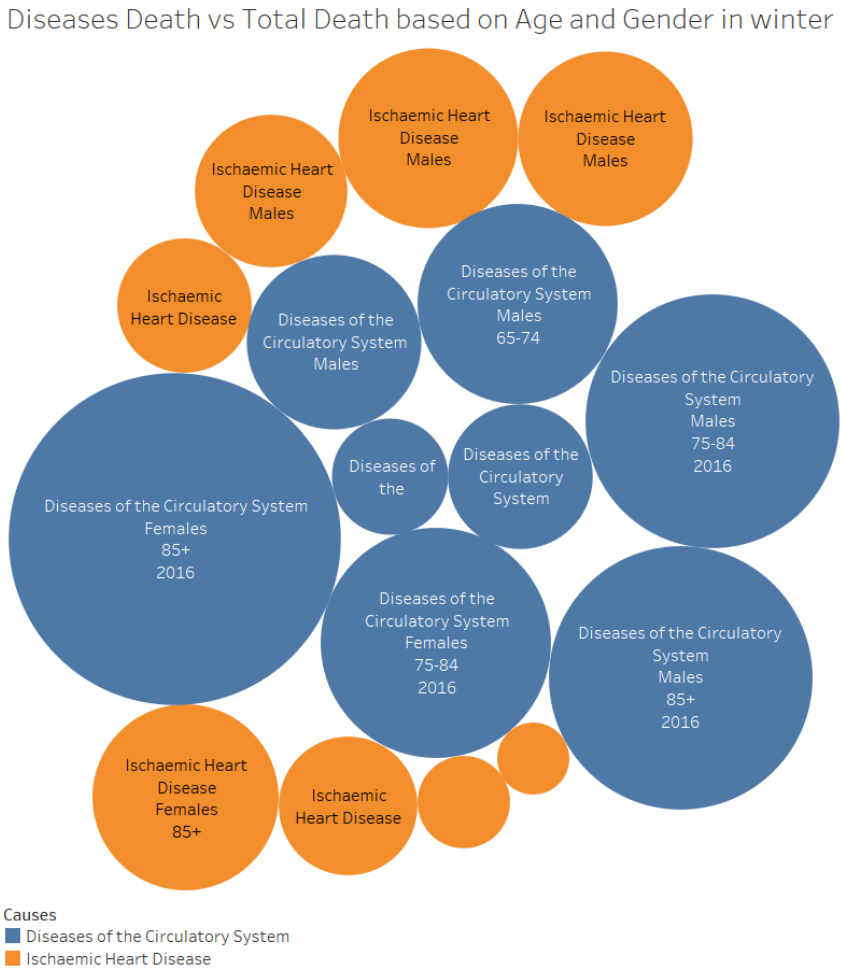

Figure 7: Winter death ratio

7.5 Discussion

This report aims to identify the total number of death related to Heart diseases based on gender and age group.By taking consideration of diseases death obesity and Overweight are major risk factor of heart failure.

According to the Public Health England, nearly two-third of adult in England are overweight, however the analysis shows that 75.9 percentage of people are obese and overweight[2].As observed previously,almost 44 % of male are overweight in the year of 2010 and ratio is slightly decreases and reach to 40% in year 2016 compare to female, in contrast up to 27% of female are obese compare to male which is 24%.In the year 2016 overweight and obesity rate are goes down.So,there is clear assumtion that people.In the year 2014 and 2015 almost 60% people have body mass index of 25 and above which represent obesity and overweight among adult.The Causes of this factor is lack of physical activity and unhealthy diet and eating habits.

People gain weight when they eat more calories which is indirectly affect the BMI. People having BMI(Body Mass Index) of 25 and above are suffering from obesity and overweight which will lead to increases the chances of heart diseases.As per the British Heart Foundation, Heart and Circulatory causes one quarter of death in England and Wales.As observed earlier,almost 4000 people died due to Cardiovascular and ischoemic diseases over the total death.

		Age Group							
		0-64		65-74		75-84		85+	
Causes	Gender	Avg. Death	Avg. Total ..	Avg. Death	Avg. Total ..	Avg. Death	Avg. Total ..	Avg. Death	Avg. Total ..
Diseases of the	Females	381	56,778,167	589	56,778,167	1,702	56,778,167	3,262	56,778,167
Circulatory System	Males	957	56,778,167	1,117	56,778,167	1,919	56,995,400	1,859	56,778,167
Ischaemic Heart	Females	150	56,778,167	259	56,778,167	665	56,778,167	1,107	56,778,167
Disease	Males	583	56,778,167	669	56,778,167	1,034	56,778,167	849	56,778,167

Figure 8: Winter death ratio

8 Conclusion and Future Work

(Partially) answer your research question and discuss the implications of your (partial) answer, talk about the efficacy of your research, and discuss its limitations.

...

Present **MEANINGFUL** future work. Sweeping more parameters in your simulation / model / platform is probably not meaningful. More discuss what could a follow up research project do, to better / differently approach / extend etc. your work.

References

Beloglazov, A. & Buyya, R. (2015), 'Openstack neat: a framework for dynamic and energy-efficient consolidation of virtual machines in openstack clouds', *Concurrency and Computation: Practice and Experience* **27**(5), 1310–1333.

Feng, G. & Buyya, R. (2016), 'Maximum revenue-oriented resource allocation in cloud', *IJGUC* **7**(1), 12–21.

Gomes, D. G., Calheiros, R. N. & Tolosana-Calasanz, R. (2015), 'Introduction to the special issue on cloud computing: Recent developments and challenging issues', *Computers & Electrical Engineering* **42**, 31–32.

Hall, M., Mazarakis, A., Chorley, M. & Caton, S. (2018), 'Editorial of the special issue on following user pathways: Key contributions and future directions in cross-platform social media research'.

Kune, R., Konugurthi, P., Agarwal, A., Rao, C. R. & Buyya, R. (2016), 'The anatomy of big data computing', *Softw., Pract. Exper.* **46**(1), 79–105.

Appendix

R code example

```
#Calculate CE for each counterparty
Value.A <- data.frame() #MTM value of each contract within cp A
# ...
for(i in 1:length(foo)){
  if(isTrue(as.character(portfolio_data[i,1])=="A")==TRUE){
    # ...
  }
}
```