A Project Report on

First Step into case study (Data Analysis)

Under supervision of Prof. Dr. Ajinkya Prabhune,
Professor, SRH university

Prof. Dr. Barbara Sprick
Professor, SRH university

Prof. Dr Herbert Schuster
Professor, SRH university

Dr. Frank Schulz
SRH university

Submitted by
Medisetti Narendra, Matrikel Number: 11012094
M.Sc in Big data & Business Analytics, SRH University, Heidelberg

First Step into case study (Data Analysis)

1. Abstract:

Our aim of this project is to analyse data and predict few business questions with better explainability and visualization.

For the Funeral dataset, we use R Programming techniques to predict the death rate, total number of deaths in a year and area wise how many deaths occurred. We can predict the death occurred at different age groups. By detail data analysis we can give various conclusions for steps to be taken for decreasing the death rate.

For Titanic data, we see how we can use R Programming techniques to predict survivors of the Titanic passengers. With the use of R packages and libraries (psych, RColorBrewer, ggplot2, corrplot, DMwR, Amelia) and an available dataset we take a look at what issues or classifications of passengers have a convincing relationship towards survival for passengers. We can try to infer their relationship for the chance of survival from the disaster. We try getting some interpretation from data using various visualization techniques like bar graphs, pie charts, column charts etc.

For the Movie dataset, we use R Programming techniques and the use of R packages and libraries (psych, RColorBrewer, ggplot2, corrplot, DMwR, Amelia) to analyze data with IMDB rating, gross, budget. With all the data available we can give a clear explanation and visualization for the top movies and top directors. This analysis helps to minimise the revenue for the movie and the necessary steps can be taken for getting profits.

For the sensor dataset as there is no target variable to predict or analysis we use unsupervised learning for explaining and analysing data. We use K-Mens clustering technique as unsupervised learning. This is used when we are having unlabeled data. The goal of this algorithm is to groups in the data with similar properties and number of groups is represented by K.

2. Dataset 1: Prediction of features and in-depth analysis for Funeral data

Aim:

To find Statistical information and in-depth analysis of the given funeral data sets by data cleaning and data analysis.

Software Requirements:

- R, R Studio:R is a language and environment for statistical computing and graphics.
- MS Excel:To import and convert the data from text to columns.
- Power Bi: It's a visualization Tool, to visualize the pre processed data.
- Packages:GGPlot., Base R.

Methodologies:

- Predicting the features of the data sets
- Data pre-processing
- Binning and calculating the required features
- Data Analysis and Visualization

Approaches:

- **Predicting Column names:**

We were given with the funeral data of 11450 rows and 10 unnamed columns. We have taken assumptions and named the columns as mentioned below

Funeral_date	First_Name	Middle_Name	Last_Name	Date_Of_Birth
03-04-2012 00:00	Hannelore	Laubenstein	Leutfeldt	18-04-1932

Death_date	Pincode	Town_City	Street	Street_Building_no
01-04-2012	55283	Nierstein	Ringstraße	39

As 1st column dates varies between short duration hence we predicted as date of funeral, 2nd, 3rd,4th columns are name fields.

By investigating 4th and 5th columns seems to be Date of birth and date of death respectively. 7-10 columns are address fields. 7th column is five digit and numeric fields hence we named it as pin code. 8th, 9th, 10th columns are Town/City, Street name and Street/Building number respectively.

- **Data Pre Processing:**

1. Removing the duplicate rows - Out of 11450 rows in the given dataset, we found that 6701 rows are unique and we removed duplicate rows by using R Programming.

2. Handling Null values –as we found loads of "NA" or "Null" values in all columns we are ignoring name fields as they won't add any importance in data analysis.

   To calculate age, Date of birth and Date of death are mandatory fields and should not be null. Hence we are removing the rows which are "NA" or "Null" values either in Date of birth or Date of Death by using R Programming.

   Total rows in dataset after removing Null values: 5872.

3. Changing date format to standard R date format (yyyy-mm-dd).

4. Calculating numberof days between Date of Death and Date of funeral: By using R Programming we calculated the number of days between Date of Death and Date of Funeral and we created a new column name DOF-DOD.

5. Calculating Age: By using R Programming we calculated the age using date of Death and Date of Birth.

6. Removing Rows with age < 0: For few rows date of death is less than date of birth, so we are removing these rows with R Programming.

   Total rows in dataset after removing age < 0: 5859.

7. Binning Age: For clear data analysis we are creating 3 buckets for age as below

Age	<50	50-65	>65

8. Binning DOF-DOD: For clear data analysis we are creating 3 buckets for DOF-DOD as below

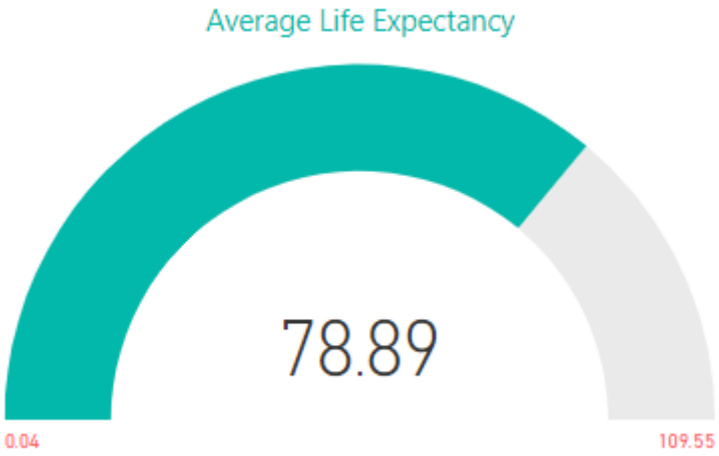
DOF-DOD	<5	5-90	>90

9. Binning Pin Code: Extracting first two digits of pin code and considering it defines a specific area, using R programming we are creating three buckets of area codes.

Pin Code	0-35	35-60	60-99

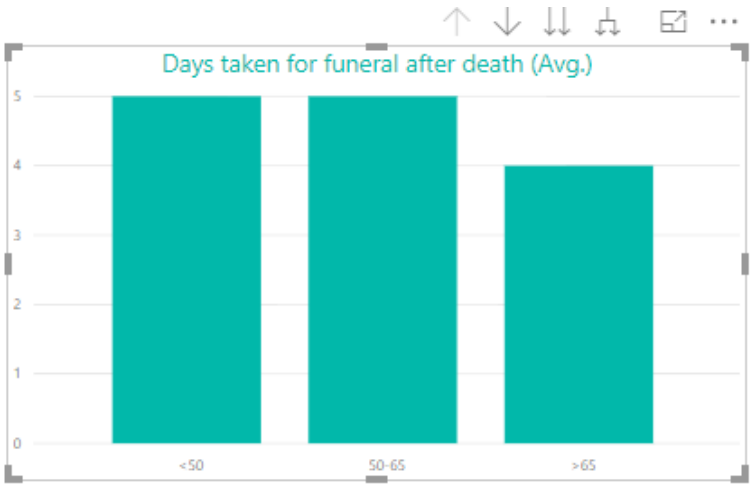
- ## Data Analysis and Visualization:

1. **Analysing Average life Expectancy:** by analysing calculated age in the given dataset the average age of a person is **78.89 years**.



By visualizing the above graph minimum age of a person is 0.04 years and the maximum age is 109.55 years. Average life Expectancy is represented in green.

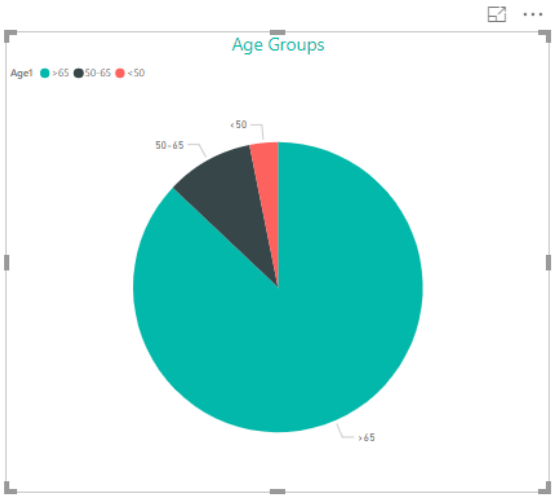
- **Days Taken for Funeral after Death(Avg.):** By analysing data from below bar graph we can conclude that



By considering average days taken to perform funeral after death on Y-axis and three buckets of age groups <50years, 50-65years and >65years on X-axis we can tell that average of 5 days for age group of <50years, average of 5 days for the age group between 50-65years and average of 4 days for the age group of >65years.

We can say that for people above age 65, the days taken to perform the funeral after death is less than the other age groups. This could be attributed to the fact that as people grow older, generally their social group size is less than people who are younger. Since more people try to pay condolences, the time to perform the funeral has increased in the case of the younger age group.

- **Age Groups:**By considering the below pie chart we get the count of the people of the different age groups so, here we take the age groups as >65, 50-65 and <50.

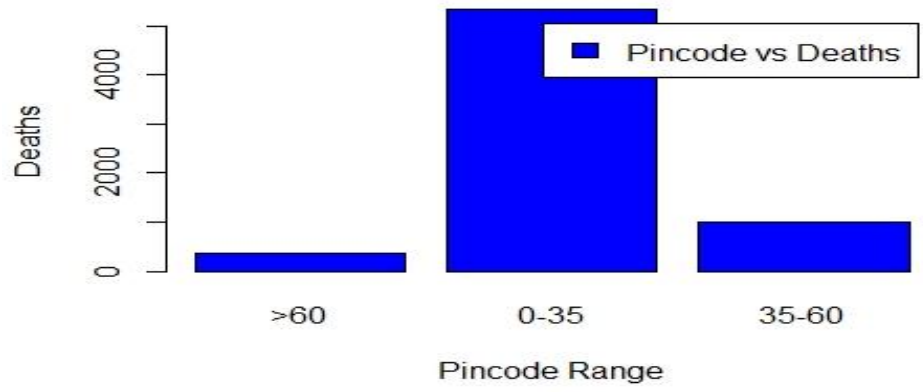


Age	Count of Age
>65	4959
50-65	558
<50	180

1. Total number of people under the age range >65: 4959.
2. Total number of people under the age range 50-65: 558.
3. Total number of people people under the age range <50: 180.

From the above graph, it can be seen clearly that the number of people dying after 65 years has increased by a substantial amount. More information regarding the causes of the death can help us to identify the steps which can be taken to prolong life. Clearly once a person reaches 50 years, he/she should be more careful about their health as the chances of dying increases sharply above 65 years of age. This could also help their families to put more emphasis on their safety and health.

Pincode vs Deaths: From the below bar graph we can get the total number of deaths in the different regions. We divide them in the three pin code ranges (>60, 0-35 and 35-60).



The above graph illustrates, the geographical area covering with pin code 0-35 has highest number of deaths . The reason might me the area has densely populated and it covers the most of the city part, hence the death rate is higher than rest, so awareness programs should to be conducted in those areas and the proper care should be taken to decrease the percentage of the deaths.

Additional data sources:

As per given data, we are unable to predict as data is not complete and had lot of missing values.

If we can add few columns like sex, cause of death would be helpful to predict target feature.

Summary:

From the funeral data set here we conclude that, the number of people dying after 65 years has increased by a substantial amount. More information regarding the causes of the death can help us to identify the steps which can be taken to prolong life. Clearly once a person reaches 50 years, he/she should be more careful about their health as the chances of dying increases sharply above 65 years of age. This could also help their families to put more emphasis on their safety and health. This data analysis is useful for the government and it helps for them to have a idea abouth deaths in area wise and gender wise. It is more useful or funeral directors and people who want to plan funerals. So proper awareness should be created anong the people regarding there health condition.

3.Dataset-2: Predicting survival and analysis of Titanic survival data

Aim:

To find Statistical information and in-depth analysis of the given Titanic passengers list data sets by data cleaning and data analysis.

Software Requirements:

- R, R Studio:R is a language and environment for statistical computing and graphics.
- MS Excel: To import and convert the data from text to columns.
- Power Bi: It's a visualization Tool, to visualize the pre processed data.
- Packages:GGPlot., Base R.

Methodologies:

- Predicting the features of the data sets
- Data pre-processing
- Binning and calculating the required features
- Data Analysis and Visualization

Approaches:

- **Overall Data Quality:**

For the project we used Titanic dataset which was provided by SRH Heidelberg University.We were given with the Titanic data of 950 rows and 12 columns are mentioned below.

Passenger ID	Survived	PClass	FamilyName, givenName	Sex	Age	SibSp	ParCh	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	Male	22	1	0	A/5 21171	7.25		S

The samples for our project and their related labels of the passenger,whether survived or not. For each passenger, given a feature names passenger ID, P class, name, sex, age, number of siblings/spouses aboard, number of parents/children aboard, ticket number, fare, and cabin embarked. The titanic dataset is not complete, meaning that for several columns, one or many of fields were not available and marked empty (especially in the latter fields – age, fare, and cabin). However, all sample points contains at least information about name, passenger id, gender and passenger class, embarked. To regularize the data, we replace missing values with KNN Imputation of those columns.

- **Data Pre Processing:**

1. Firstly, we have to analyse feature and missing values or NULL in each column, we found that age and cabin column has NULL values by using R Programming.

2. **Analyse which features in data can analyze survival of passenger**.

Hence, we are removing the columns which are not mandatory fields and don't add any value in analysing target feature (Survival), we are removing PassengerId,Name, Ticket, Cabin, embarked by using R Programming.

Total columns in dataset after removing columns fields: 7

3. **Calculating age**: The missing value or NULL in age column can be added by using KNN Imputation method by using R Programming.
4. **Binning age**: For clear data analysis we are creating 3 buckets for age as below

Age	<25	25-45	>45

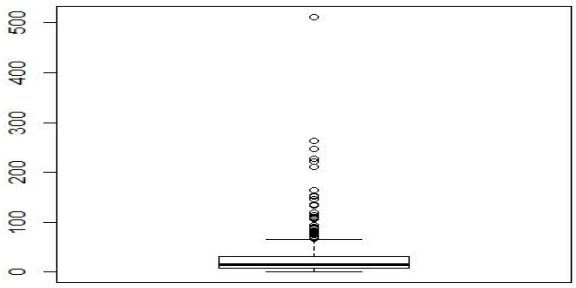
5. **Binning sibsp**: For clear data analysis we are creating 3 bucket for sibsp as below

Sibsp	0	1	>=2

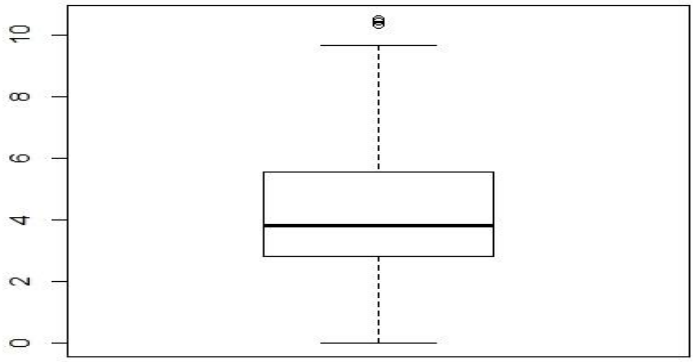
6. **Binning Parch**: For clear data analysis we are creating 3 buckets for Parch below

Parch	0	1	>=2

7. **Removing outliers for Fare**:We can observe few fare values as outliers in below box plot, to handle outliers we use R programming and fit all outlier data to 99th percentile of fare.



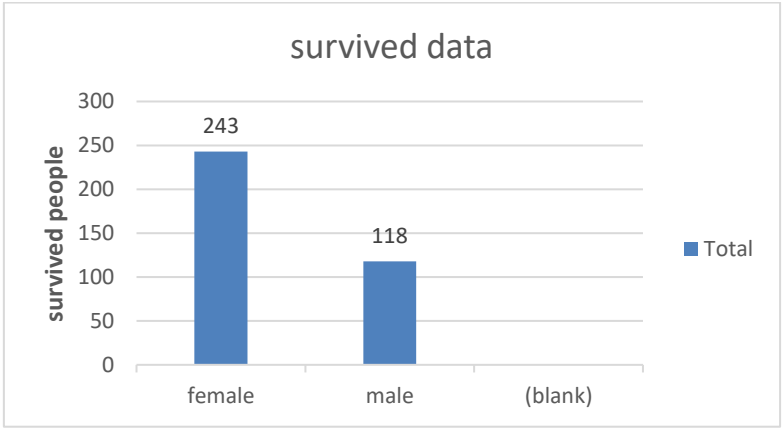
8. After removing extreme outliers in fare feature, we can visualize data in box plot as below



Data Analysis and Visualization:

1. **Analysis of male and female survived**: By analysing calculated sex in the given dataset who are survived

.



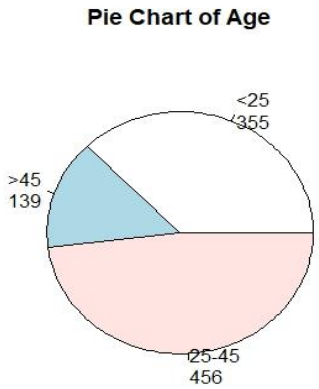
The above graph illustrates that, 243 female passengers and 118 male passengers are survived. Female passengers are first evacuated and they are given first preferences for the help somost percent of the female passengers are survived.

Calculating survived people based on sex: By using R Programming,We calculated the number of male and female passenger who is survived or dead.

Survived/Dead	Male	Female
Not Survived	493	91
Survived	118	243

By visualizing the above graph show 243 female passenger and 118 male passengers are survived.

- **Age group**: By analysing data from below bar graph we can conclude that,



Using R programming analysis we can calculate age of passenger in titanic data.

- The total number of passenger whose age is <25 are 355.
- The total number of passenger whose age is between 25-45 agesis456.
- The total number of passenger whose age >45 are 139.

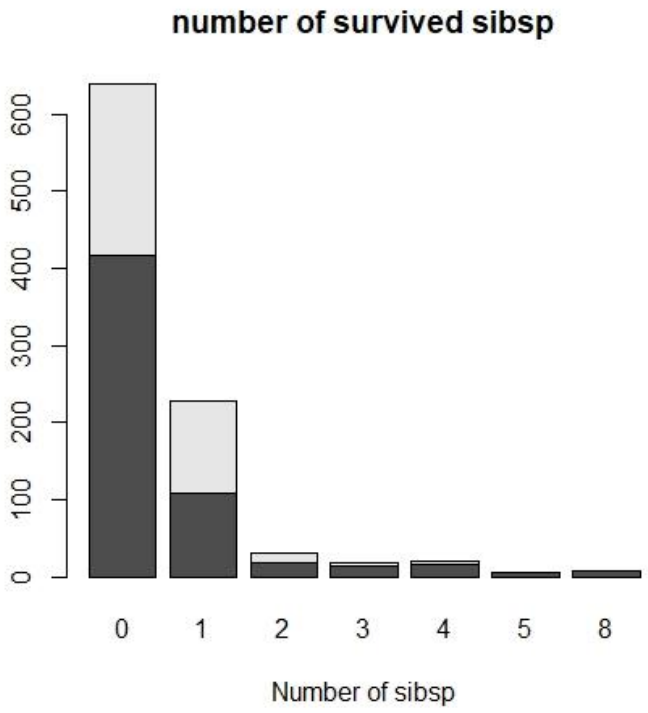
The above graph illustrates those passengers with the age less than 25 years, 37% of them are survived. Passengers with the age between 25-45 years, 48% of them are survived. Passengers with the age more than 25 years, 14% of them are survived. Here we conclude that old aged passengers are least survived and the middle aged passengers are mostly survived. Measures should taken to save the old age people more because the middle aged are strong enough to help themselves than the old aged people.

Calculating which age survived or not: By using R programming we can calculate whichage group are survived or not survived.

Survived or not	<25	25-45	>45
Not Survived	200	289	97
Survived	155	167	42

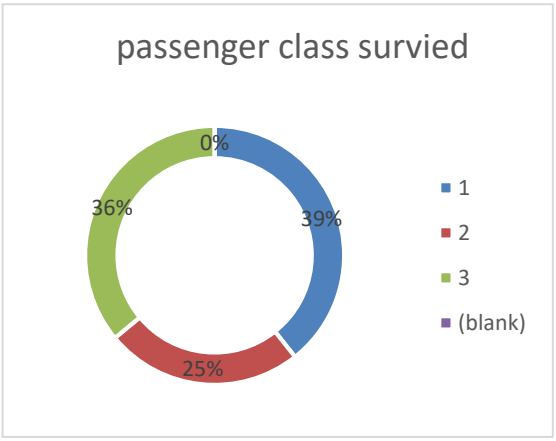
From the above table we can say that, passengers with the age less than 25 years, 155 of them are survived.Passengers with the age between 25-45 years, 289 of them are survived. Passengers with the age more than 45 years, 42 of them are survived.

Number of siblings survived: By analysing data from below bar graph we can Conclude thenumbers of siblings, who are survived in titanic data.



By considering above graph, passenger who has no siblings survived is 222 passenger and 417 are not survived. Single siblings passenger survived are 118 people and 109 passengers with single siblings are not survived. Three siblings passenger are survived are 4 and died are 13 passengers. Passenger, who had 5 and 8 siblingsis no passenger is survived.

- **Passenger class:**Calculating passenger survived in three different class



Pclass	No. of Survived
1	39%
2	25%
3	36%
Grand Total	**100%**

By conceding above graph we can say, first passenger class survived more with 39% Second passenger class with 25% and third class with 36%. By conceding above graph can say first class was more safety than other classes.

From the above donut chart we conclude that more first class passengers have survived and less third class passengers have survived. There should be same safety for all class passengers. More number of life jackets are should be carried and all the proper safety measure should be taken.

Summary:

From the titanic passengers list data set here we conclude that, majorityof passengers with first class tickets who are children, female are mostly survived. Everyone cannot be survived but the safety measures should be increased like there should be proper awareness for the safety in the ship and there should be proper demonstration for the evacuation when there is any accident. More number of the life jackets and lift bots should be used.

4. Dataset-3: Predicting movie trend and analyzing data

Aim:

To find Statistical information and in-depth analysis of the given movie data sets by data cleaning and data analysis.

Software Requirements:

- R, R Studio :R is a language and environment for statistical computing and graphics.
- MS Excel : To import and convert the data from text to columns.
- Power Bi : It's a visualization Tool, to visualize the pre processed data.
- Packages:GGPlot, Base R.

Methodologies:

- Predicting the features of the data sets
- Data pre-processing
- Binning and calculating the required features
- Data Analysis and Visualization

Approaches:

Dealing with missing data:

For this project we used Movie dataset which was provided by SRH Heidelberg University. We were given Movie data of 5033 rows and 29 columns.The samples for our project and their related characterize to deals with missing data, wrong data and understand data with unclear semantics. Analysis movie data with identifying top movie, gross, high budget, low budget, number of movies, top director and relation between variables.

Data Pre Processing:

1. Imputing missing data of each variable in movie data. Handling wrong data and make clear data with unclear semantic in columns.

2. In movie data removing rows which are having NULL or missing value can be imported using KNN imputation (nearest neighbour).

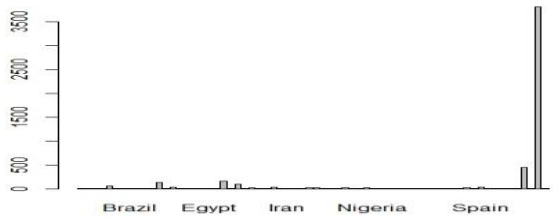
3. Genres been identified with first phase of data, which explains movie's primary genre in movie data by using R programming.

4. Aspect ratio been given as Month-value, where we can extract only value for aspect ratio using substr function in R programming

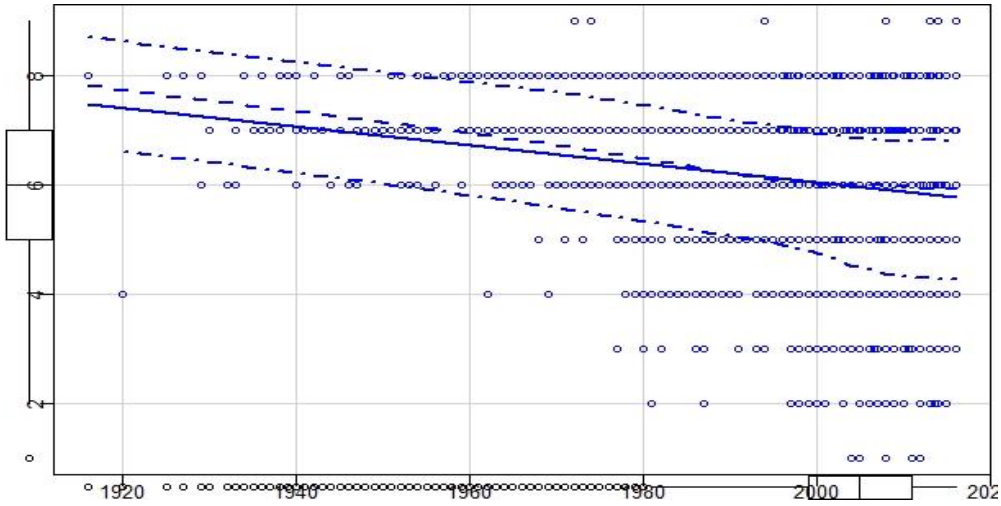
Data Analysis and Visualization:

Explore movies by year: Data provided is movies within duration of 1916-2016. Highest numbers of movie release in a year 2009 are 260 movies and 2014 are 258.

95% of movies are from UK and USA as per below data analysis bar plot

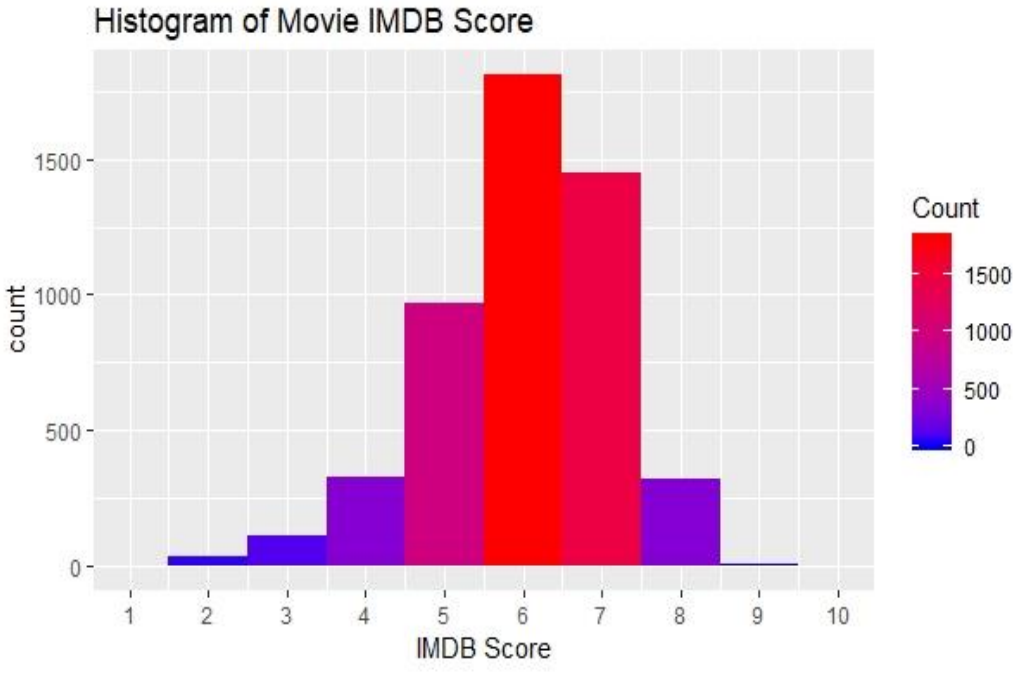


Analysis of data for year vs IMDB scores:



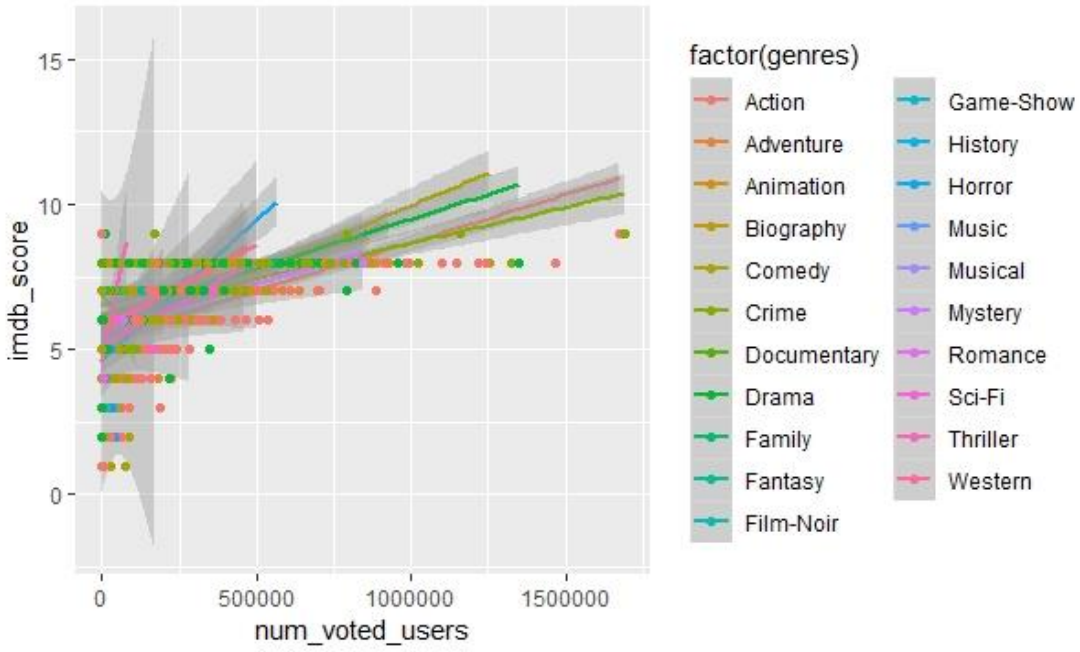
By visualizing above scatter plot, most of movies having imdb rating >8 are between 2000-2014

Analysis between IMDB score and count of movies:



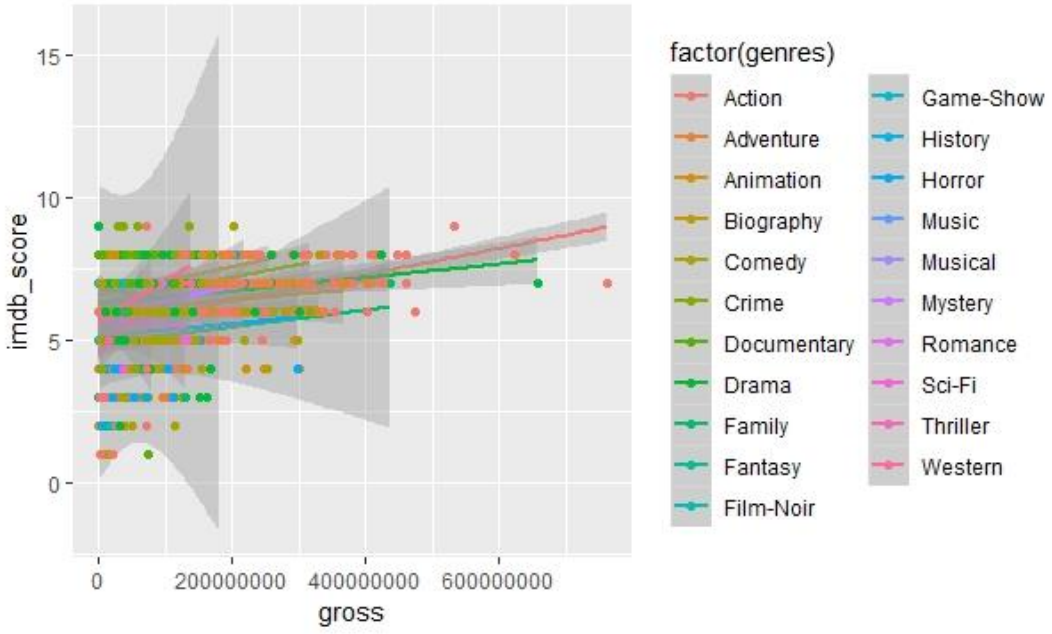
As per above histogram, most of movies in given data set had imdb score 6 and 7.

Analysis of data imdb score vs. number of voted users vs. Genres:



As per above graph, we can analyse that most users voted for Action movies and highest imdb rating is for crime, documentary and comedy movies. We can predict that most users in imdb are likely to watch Sci-Fi, Thriller and crime/comedy movies.

Analysis between Gross revenue vs. IMDB score:



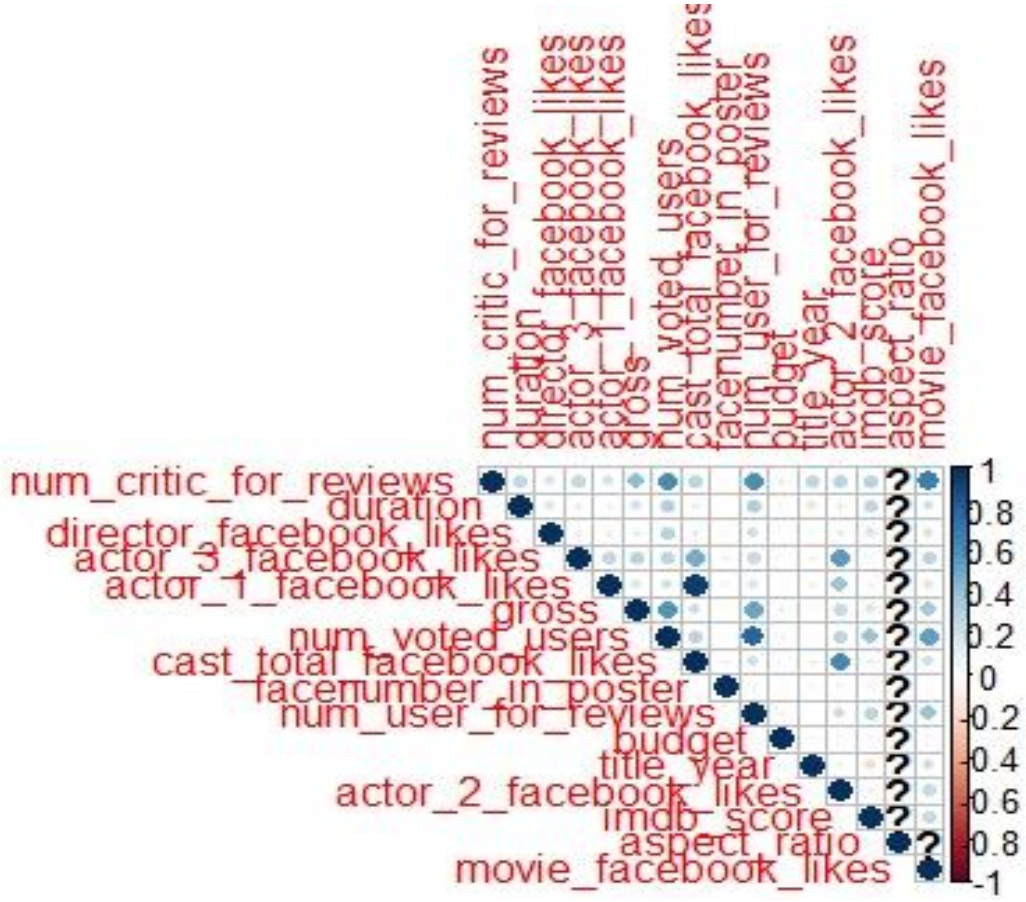
From above plot, we can analyse that Action, crime movies are most likely to collect highest gross and had highest IMDB rating.

Gross revenue is highly correlated with IMDB rating positively

Correlation plot for all variables in movie data set:

Correlation between all variables in Movie data has been calculated and a graph has been plotted for it as mentioned below.

Correlation plot shows correlation between features as highly correlated with darker in colour and weakly correlated as lightest colour in blue.



From above correlation plot, we can see actor facebook likes and cast total likes are highly related, Number of reviews and facebook likes are highly related, duration and gross are highly related to critic reviews.

1. **Top ten movies:**By using R programming we can calculate top 10 movies with highest Gross, highest budget, lowest budget in movie data.

Top 10 movies	Highest gross	Highest budget	Lowest budget
1	Avatar	The Host	Tarnation

2	Titanic	Lady Vengeance	My Date with Drew
3	The Avengers	Fateless	A Plague So Pleasant
4	The Avengers¬†	Princess Mononokey	The Mongol King
5	The Dark Knight	Steamboy	Clean
6	Star Wars: Episode I - The Phantom Menace	Akira	El Mariachi
7	Star Wars: Episode I - The Phantom Menace	Godzilla 2000	Primer
8	Avengers: Age of Ultron¬†	Kabhi AlvidaNaaKehna	Cavite
9	The Dark Knight Rises¬†	Tango	Newlyweds

Top 10 directors:by using R programming we can calculate top 10 movie directors number of movies and highest gross revenues in movie data.

Top 10 director Number of movies	Top 10 director Highest gross revenue
Steven Spielberg 26	James Cameron
Woody Allen 22	James Cameron
Clint Eastwood 20	Joss Whedon
Martin Scorsese 20	Joss Whedon
Ridley Scott 17	Christopher Nolan
Spike Lee 16	George Lucas

Steven Soderbergh 16	George Lucas
Tim Burton 16	Joss Whedon
RennyHarlin 15	christopher Nolan
Oliver Stone 14	Nolan Andrew Adamson

Summary:

As per our analysis, Movies with highest IMDB rating, User reviews, Movies with genre Action/Comedy/Documentary, duration plays a major role in gross collections of a movie.

5. Dataset-4: Unsupervised learning Sensor data

Title:

To find Statistical information and in-depth analysis of the given senor data sets by data cleaning and data analysis.

Software Requirements:

- R, R Studio:R is a language and environment for statistical computing and graphics.
- MS Excel: To import and convert the data from text to columns..
- Packages:GGPlot, Base R.

Methodologies:

- Predicting the features of the data sets
- Data pre-processing
- Unsupervised learning, clustering
- Data Analysis and Visualization

Approaches:

Column names assumption: As column names were not mentioned in sensor data, we assumed first column in given data set is number of sensor and their respective reading in row wise.

As we are considering sensor numbers are features for given data set, we are transposing data rows to columns using excel functionalities.

Unsupervised learning:

As we don't have any specified target variable in given data set, we can do unsupervised learning for understanding and analysing data

Data Analysis and Pre processing:

Data pre processing:

In a given data set for sensor data, we have 317 rows and 547 columns. As per our assumption rows gave number of sensors and columns gives study of sensors for duration with specific time.

As we couldn't take features as rows, we are transposing data and considering each feature as a sensor with number from 1 to 317.

Data Analysis:

For given data set, as we don't have any specific target variable to predict we go with unsupervised learning for learning pattern or properties of given data.

We have chosen cluster approach which can clearly represent data and its properties with better learning.

K-Means Clustering:

K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable *K*. The algorithm works iteratively to assign each data point to one of *K* groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the *K*-means clustering algorithm are:

1. The centroids of the *K* clusters, which can be used to label new data
2. Labels for the training data (each data point is assigned to a single cluster)

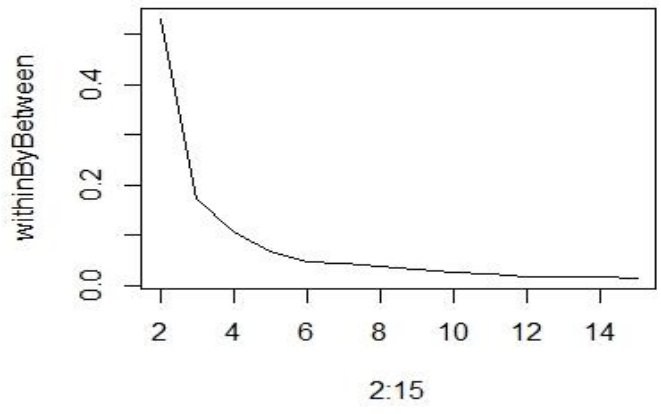
Rather than defining groups before looking at the data, clustering allows you to find and analyse the groups that have formed organically. The "Choosing K" section below describes how the number of groups can be determined.

Each centroid of a cluster is a collection of feature values which define the resulting groups. Examining the centroid feature weights can be used to qualitatively interpret what kind of group each cluster represents.

Clustering Sensor data:

We created clusters using k means cluster algorithm using R programming for sensor data.

First we created 10 clusters primarily and calculated their centroids, betweenSS and withSS values to find better explain ability and plotted graph between and within SS to find ideal number of clusters to analyse data.



By observing above graph Between SS and number of clusters, explainability on cluster after 6 seems to be decreased.

So we are taking 6 clusters and below table explains each cluster by aggregating data with similar properties into cluster.

Cluster/Sensor Node:	X1	X2	X3	X4	X5	X6
1	238.1689	324.1816	2323.035	773.7683	2239.929	118.8271
2	282.6342	654.1289	3237.742	1058.673	3833.144	97.54531
3	215.0876	587.044	2541.38	1147.357	4687.641	158.8271
4	196.3006	201.1087	2035.176	840.7631	1687.88	36.82155
5	328.6682	1066.409	4301.29	1361.722	6564.743	227.1016
6	330.1355	809.9606	3895.825	1247.271	5667.725	223.1943

By observing above sample clustering table, we can explain probability of each cluster by its metrics

Cluster and Number of rows explained:

Clusters	**1**	**2**	**3**	**4**	**5**	**6**
No of rows explained	19	97	54	92	245	40

Summary:

We have identified features of sensor data and analysed its pattern by unsupervised using K means clustering

6. Bibliography:

1) https://www.datascience.com/blog/k-means-clustering
2) http://rstatistics.net/r-lang-practice-exercises-level-1-beginners/
3) http://www.sthda.com/english/wiki/visualize-correlation-matrix-using-correlogram
4) https://r4ds.had.co.nz/data-visualisation.html
5) http://rstatistics.net/r-lang-practice-exercises-level-1-beginners/
6) https://www.analyticsvidhya.com/learning-paths-data-science-business-analytics-business-intelligence-big-data/learning-path-r-data-science/