A Project Dairy for

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

Project dairy for Funerals Dataset:

Day 1:

First I analyse the dataset which has been given. There were three columns which had names, which I assumed to be so after looking at many rows. Generally, in any forms (passport or some government id forms), there are three fields – first, middle and last name. So I assumed this to be the case here. I couldn't understand what one particular column meant (column 8) but I could see that some values were repeated several times. I put it aside and checked column 9. Many values in this column had the word 'Strahe' which I knew meant street. The last column had random numbers without any pattern so I assumed that it was the building number, column 9 as the street and column 8 as the city which would give us information about the person's address. There were several columns which had dates in them, 3 to be exact. In one column, the year mentioned was about 40 to 50 years back, in the past. The other two columns had year closer to each other (in most or all cases, in the same year). Since the dataset was referring to funerals, I assumed that the date column having years in the past was the date of birth and the dates having same year to be date of death and date of funeral. Since the funeral will obviously take place after death, I assumed column names correspondingly.

Day 2 and 3:

Data cleansing:

The next step is data cleansing as some of the value for the columns were null. For cleansing the data, I deliberated using SQL, Python and R-Studio. I ruled out SQL even though it would be easy to do so using it as I felt it would get complicated with each extra layer of cleansing. I choose R-Studio simple because I wanted to practice using it as I was learning it.

After looking through the values, the following stages were identified:

Removing Duplicate:

I used the functions unique and duplicate. To remove the duplicates directly I used duplicate method. Before this, I tried to write a script for removing duplicates manually by iterating through the rows and finding distinct ones, but it was getting too complicated. I found the function to be simple and it could do the job in just a few lines of code compared to the later approach.

Removing null values:

In funeral data as we found lots of null values and empty values in the fields. To remove them I have used command as.na. In this case also, we can manually iterate through the rows and remove the ones having null and empty values but I used a predefined function as it was simpler.

Day 4:

Changing Date format to yyyy-mm-dd and adding new columns:

Funeral data has date format was incorrect which as yy-mm-dd and time in three columns. It is difficult to identify each of value in funeral data because of the timestamps. I removed it in each column using the function strptime.

Removing Dates columns which were in incorrect format:

I removed old column which had yyyy-mm-dd-ss by keeping column as Null. It reduces confusion of many variables which as same names.

Removing Rows with age < 0*:*

For few rows, the date of death is less than date of birth, so we are modifying rows by keeping age as > 0. So if date of birth is less than date of death, then these rows are removed with R Programming.

Binning Age:

To know that how many number of people whose age group less than fifty, 50 to 65 age group and then greater than 65 age group, we used if else method.

Age	<50	50-65	>65
	180	558	4960

.

Calculating number of days between Date of Death and Date of funeral:

Now we must know the estimation of in how many days has the funeral been performed after their death. To find out I searched in Google I get relevant answer in quik-r, stack overflow I pick both ideas and I executed in R studio. I used function as.Date(as.character(funeral_unique$DOF), format="%Y-%m-%d")-as.Date(as.character(funeral_unique$DOD), format="%Y-%m-%d"). A new column name DOF-DOD has been used.

DOF-DOD	<5	5-90	>90
	1438	4218	42

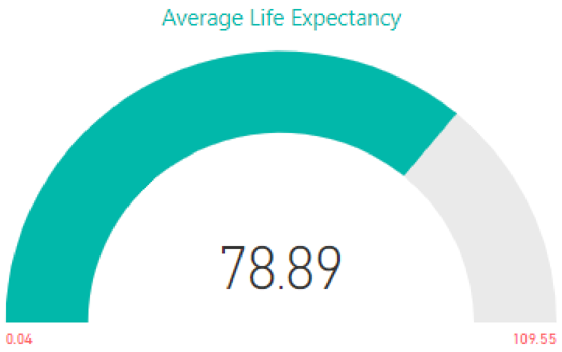
Day 5:

Pin Code:

This is to find out people who died in different regions and to know approximate number people who died. first two digits of pin code and considering it defines a specific area we created three bucket of area code 0 to 35, 35 to 60 and60 to 99.Extracting, using R programming.

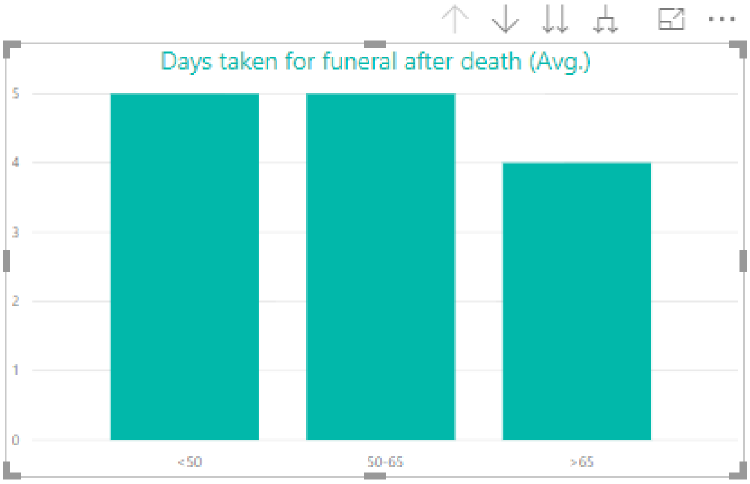
Pin Code	0-35	35-60	60-99
	326	1006	272

Analysing Average life Expectancy:



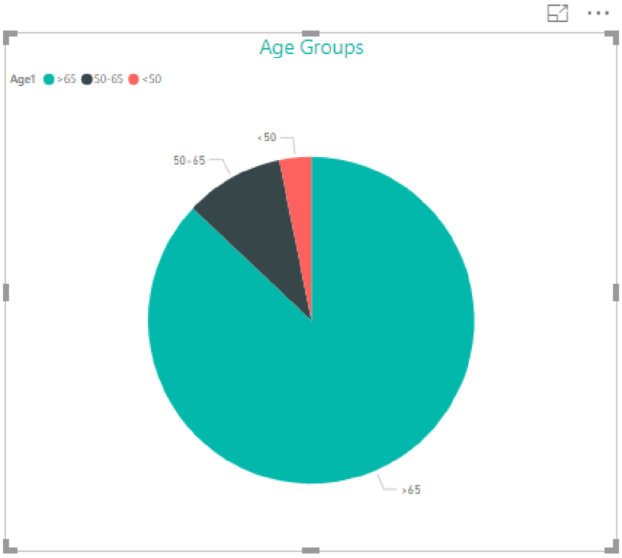
The visualization of above graph was done using Power BI. I started using Power bi it as it has good visualization tools. I also wanted to learn it even though I was used to working in Microsoft Excel. Here we are analysing average life expectancy.

Days Taken for Funeral after Death (Avg.):



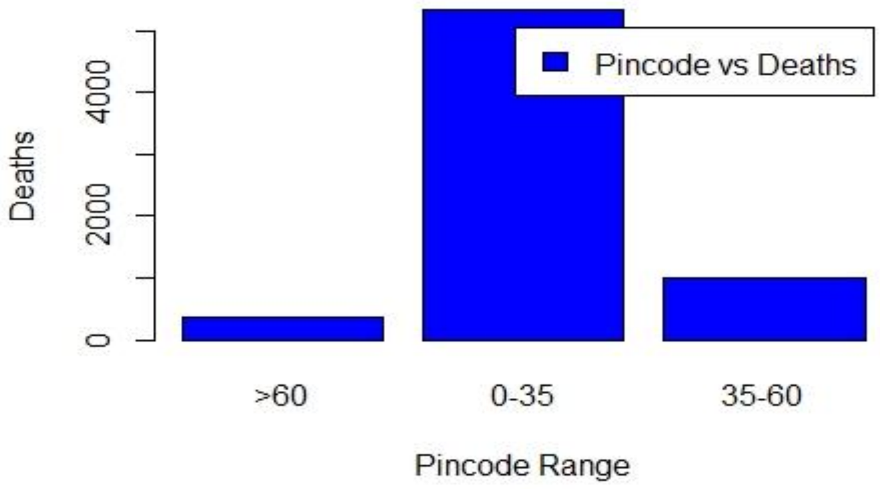
The visualization of above graph was done using power BI .By considering average days taken to perform funeral after death on Y-axis and three buckets of age groups <50years, 50-65years and >65years on X-axis we can tell that average of 5 days for age group of <50years, average of 5 days for the age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

Age Groups:



The visualization of above graph was done using power bi.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

Pin code vs death:



For the visualization of above graph, I used R studio. I Google searched the way to draw the bar graph in r programming. I also saw some comments in stack overflow about how ggplot2 package is used. I installed ggplot2 package in R studio and used barplot function. I try to use that bar plot command but I failed many times. Finally, I executed this code successfully barplot(table(funeral_unique$Pincode),col="blue", legend.text = "Pincode vs Deaths",axisnames = TRUE, xlab = "Pincode Range", ylab = "Deaths"). In the 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).

References:

- https://community.powerbi.com/t5/Community-Blog/Fun-with-Graphing-in-Power-BI-Part-1/ba-p/358399
- https://stackoverflow.com
- https://www.tutorialspoint.com/power_bi/power_bi_visualization_options.htm
- https://www.edureka.co/blog/power-bi-dashboard/
- https://exceleratorbi.com.au/extending-power-bi-r-visuals/
- https://ggplot2.tidyverse.org/
- http://r-statistics.co/ggplot2-Tutorial-With-R.html
- https://datacarpentry.org/R-ecology-lesson/04-visualization-ggplot2.html
- https://docs.microsoft.com/en-us/power-bi/
- https://support.rstudio.com/hc/en-us/categories/200035113-Documentation
- You tube videos for Power BI

Week-2 (Titanic project diary)

Day 1:

First I analysed that dataset which has been given. We were given with data related to the Titanic, of 950 rows and 12 columns. We were given column names like passenger ID, P class, name, sex, age, number of siblings/spouses aboard, number of parents/children aboard, ticket number, fare, and cabin embarked. In one of the columns, the distinct data was either 1 or 0. I assumed this to be referring to whether the passenger has survived or not. All sample points contains at least information about name, passenger id, gender, passenger class and cabin embarked. In general, I thought that we should compare each and every variables with survival column so we can find how many passengers are survived and make any prediction or observations on how many have survived for example survival rate for males vs. females, the number of people who have survived as per the passenger class, survival rates and death rates between siblings etc.

Day 2 and 3:

From the above data, I could see that many records had null values. I also had to organise and prepare data indirectly from the data which we were given. For this purpose, I decided to use R Studio simply on the account that I was learning to program in R Studio and I thought that this would be a good practice. I thought that I could do the programming in Python as well, which I believe wouldn't be difficult but I stuck to R Studio as I wanted to get a grip on it. The next step was to cleanse the data.

Data cleansing:

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 command view() in R programming. 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 Passenger Id, Name, Ticket, Cabin, embarked by giving as NULL in R programming.

Calculating age:

To calculate the age we have fill the missing value or NULL in age column can be added by using KNN Imputation method. So I search in Google which package should be used to run kNN Imputation I found DMwR used a code knn Imputation(data = data1,k=3). Then I install and run in R studio.

Day 4:

I continued to extrapolate data from the given data.

Binning age:

To know approximate how many number of people whose age group in titanic data. For clear data analysis we are creating 3 buckets for age are less than 25 age, 25-45 age group and passenger whose age is more than 45 years. So I used commend ifelse and table command. I had already used this command in funeral data and I felt this method is most the efficient to perform this particular kind of operation.

Age	<25	25-45	>45
	355	456	355

Binning sibsp:

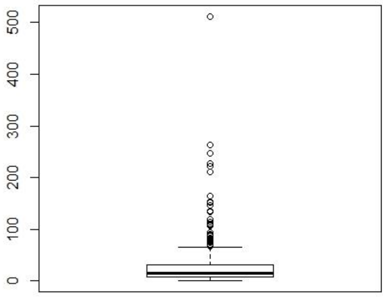
To know that how many number of sibling are survived in titanic data. For clear data analysis we are creating 3 bucket for sibsp who as no sibling, passenger who has single sibling, passenger who has more than two in titanic data.

Sibsp	0	1	>=2
	643	227	80

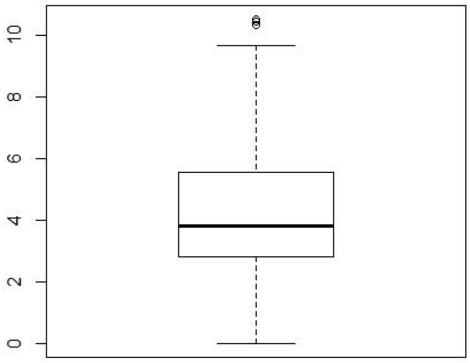
1. **Binning Parch:** To estimate number of parch in titanic data. For clear data analysis we are creating 3 buckets for Parch bellow

Parch	0	1	>=2
	724	126	100

1. **Removing outliers for Fare:**
   To outlier data of 99th percentile of fare I searched in google and I found useful information in quick-R. I used a code which is quantile(data2$fare,0.99) and I used boxplot. 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.



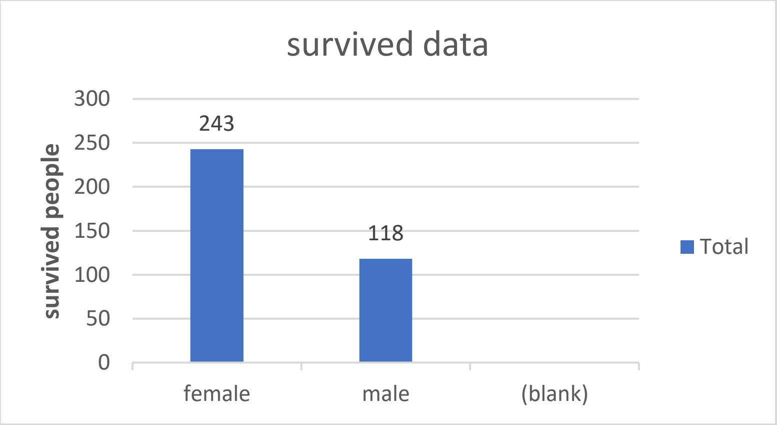
1. After removing extreme outliers in fare feature, we can visualize data in box plot as below



Day 5:

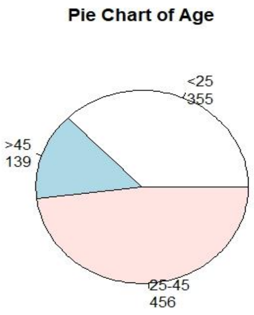
I spent some time on the visualization part as I tried to work with Power BI, but found it a bit complicated. For the visualization of below graph, I used Excel. I started using Microsoft Excel as I found it very user friendly and the fact that it has good visualization tools in graphs but most importantly because it is very simple to use. I did not continue with Power bi for visualization as I was still in the learning stages. So I choose to do in Excel because I have quite a few years of experience. The below graph tells whether female have survived more than male passengers in titanic data.

Analysis of male and female survived:



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

Age group:



For the above visualization, I used R studio, as I was trying to get a hang of the graph operations in the same. The total number of passengers whose age is <25 are 355. The total number of passengers whose age is between   25-45 ages is 456.The total number of passenger whose age >45 are 139.

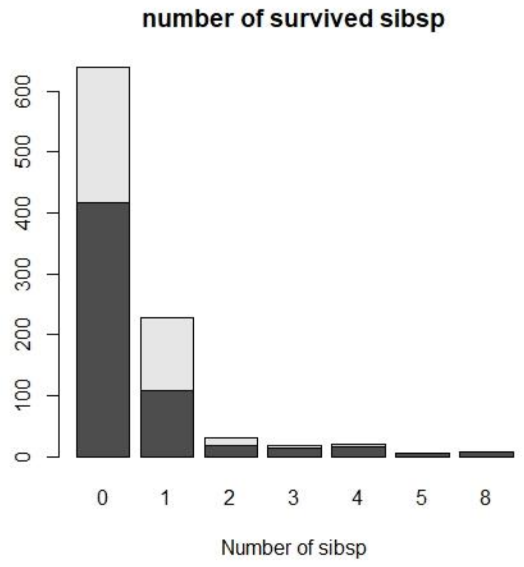
Day 6:

Calculating which age survived or not:

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

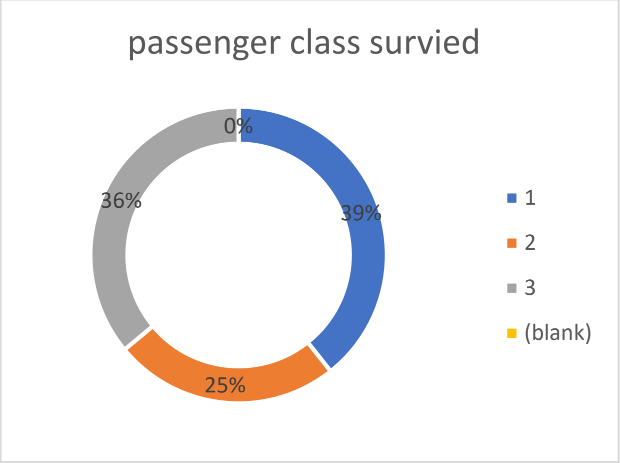
I tried to find out the survival rate with respect to age group. From the above table we can say that, in the range of passengers with the age less than 25 years, 155 of them are survived. For passengers with the age between 25-45 years, 289 of them are survived and for passengers with the age more than 45 years, 42 of them have survived.

Number of siblings survived



The above graph has been done using Excel. From the above graph, passengers who have no siblings, the number of such passengers who have survived is 222 passenger and 417 are not survived. Single siblings passengers who have survived are 118 people and 109 passengers with single siblings have not survived. Three siblings passengers who have survived are 4 and the ones who perished are 13 passengers. For passengers who had 5 and 8 siblings, none have survived.

Passenger class:



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

The above graph is done using Excel. From the above graph we can say, first passenger class survived more with 39% Second passenger class with 25% and third class with 36%. From the above graph can say generally first class had more safety than other classes.

References:

- https://www.statmethods.net/graphs/bar.html
- https://www.tutorialspoint.com/r.htm
- https://www.guru99.com/r-bar-chart-histogram.html
- https://www.excel-easy.com/examples/bar-chart.html
- https://www.smartsheet.com/bar-charting-excel-bar-graph
- https://www.statmethods.net/graphs/pie.html
- https://pythonprogramming.net/
- https://pythonspot.com/
- http://www.sthda.com/english/wiki/ggplot2-pie-chart-quick-start-guide-r-software-and-data-visualization
- https://support.rstudio.com/hc/en-us/categories/200035113-Documentation
- Youtube videos on R-studio graphs
- Microsoft Help for Excel

Movie data diary:

Week 3:

Day 1:

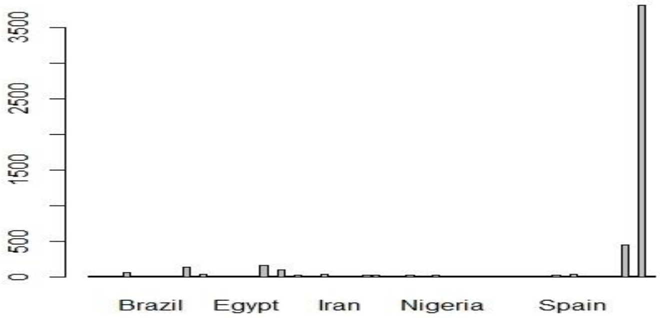
Steps of learning;

First I analyse that dataset is used in which programming. We were given with the Movie data of 5033 rows and 29 columns. I assumed to be so after looking at many columns that to find out for each variables having 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 Correlation for all variables in movie data.

Data cleansing:

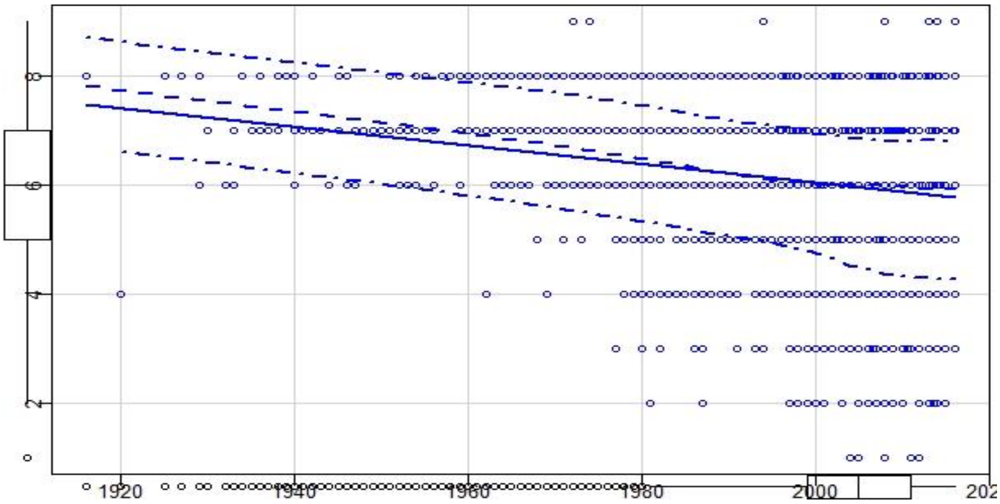
Firstly, we have to analyse variables and Imputing missing data of each variable in movie data. Handling wrong data and make clear data with unclear semantic in columns. Import Excel in R studio and view the data by using view command. find out assume which row has NULL in movie data. In movie data removing rows which are having NULL or missing value can be imported using KNN imputation (nearest neighbour). So do this KNN Imputation we have install (DMwR)library packages and run this command (movie_data<-knnImputation(movie_data, k=5) in R studio.Genres been identified with first phase of data, which explains movie's primary genre in movie data by using R programming. Aspect ratio been given as Month-value, where we can extract only value for aspect ratio using substr function in R programming.

Visualization :



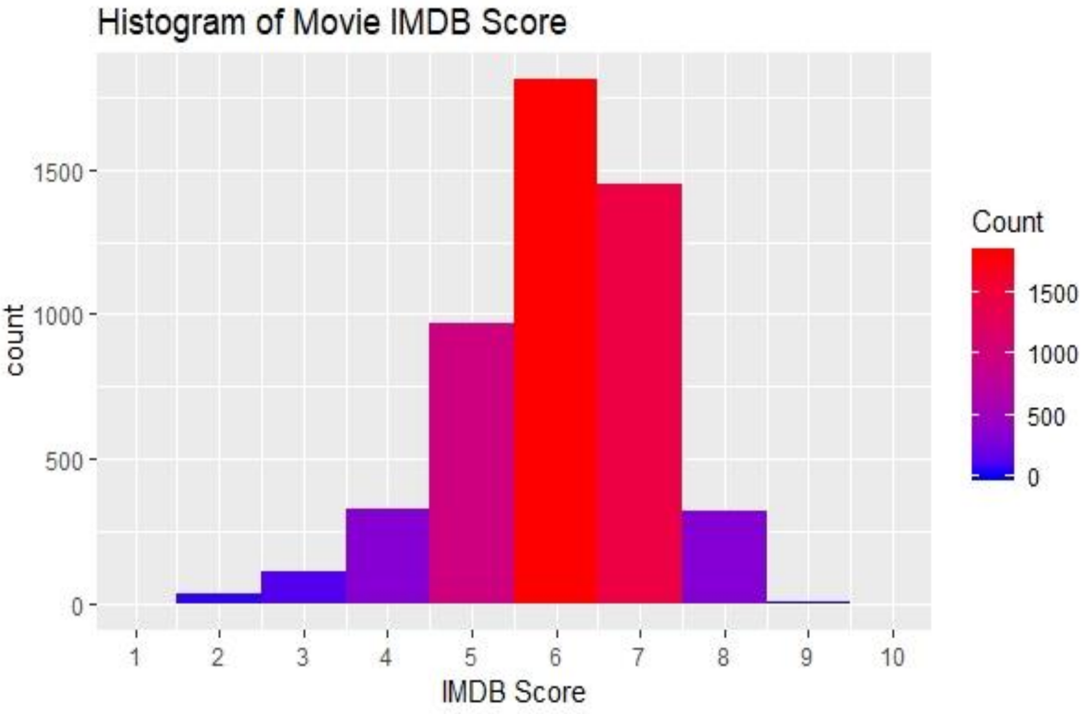
The visualization of above graph I used in R studio. Now I install ggplot2 library. Explore movies by year by using code range in R studio: Data provided is movies within duration of 1916-2016. To know Highest number of movie release in a year 2009 are 260 movies and 2014 are 258 by using command sum(with(movie_data,title_year=='2009')) and sum(with(movie_data,title_year=='2014')). We can assume above graph 95% of movies are from UK and USA.

Analysis of data for year vs IMDB score:



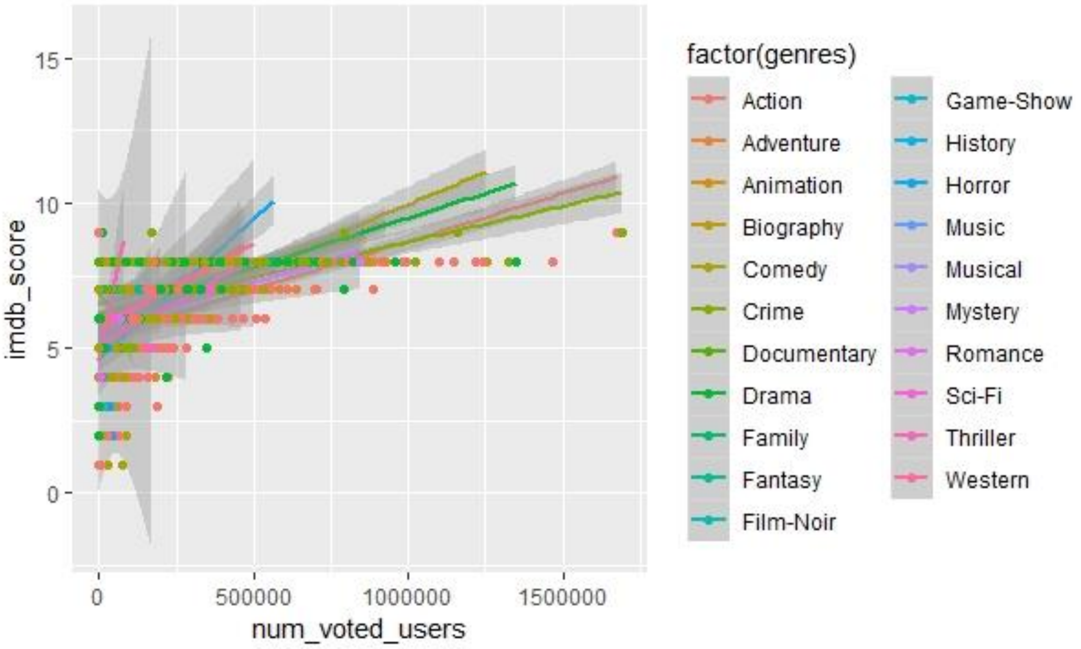
The visualization of above graph I used in R studio. First we install ggplot2 library. I search in Google for scatter plot code I found some result I put code R studio errors arrived but at finally this code scatterplot(x=movie_data$title_year,y=movie_data$imdb_score) was successful. By visualizing above scatter plot, most of movies having imdb rating >8 are between 2000-2014.

Analysis between IMDB score and count of movies:



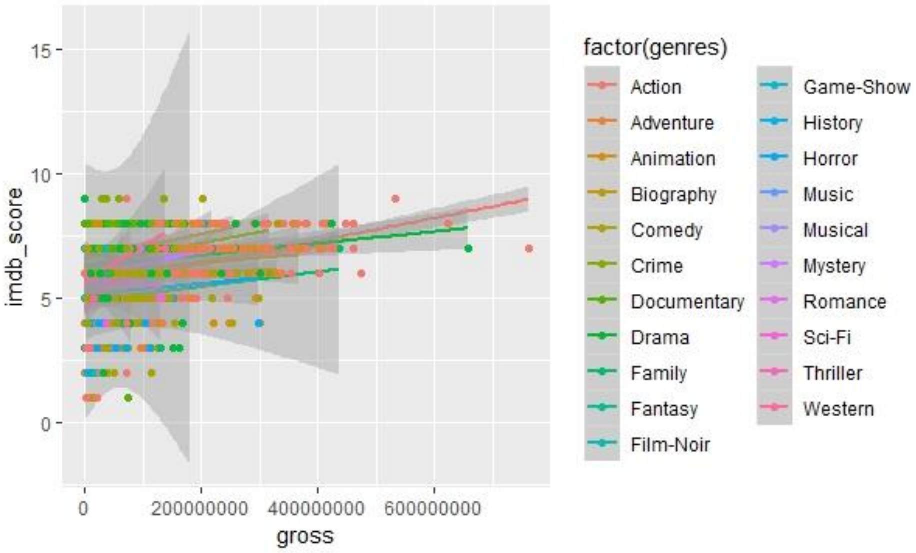
The visualization of above graph I used in R studio. The above was histogram I used code histogram.As per above histogram, most of movies in given data set had imdb score 6 and 7.

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



The above graph was done in R studio by using library(ggplot2) the command we used is ggplot(data=movie_data,aes(x=num_voted_users,y=imdb_score,colour=factor(genres)))+stat_smooth (method=lm,fullrange = FALSE)+geom_point() .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:

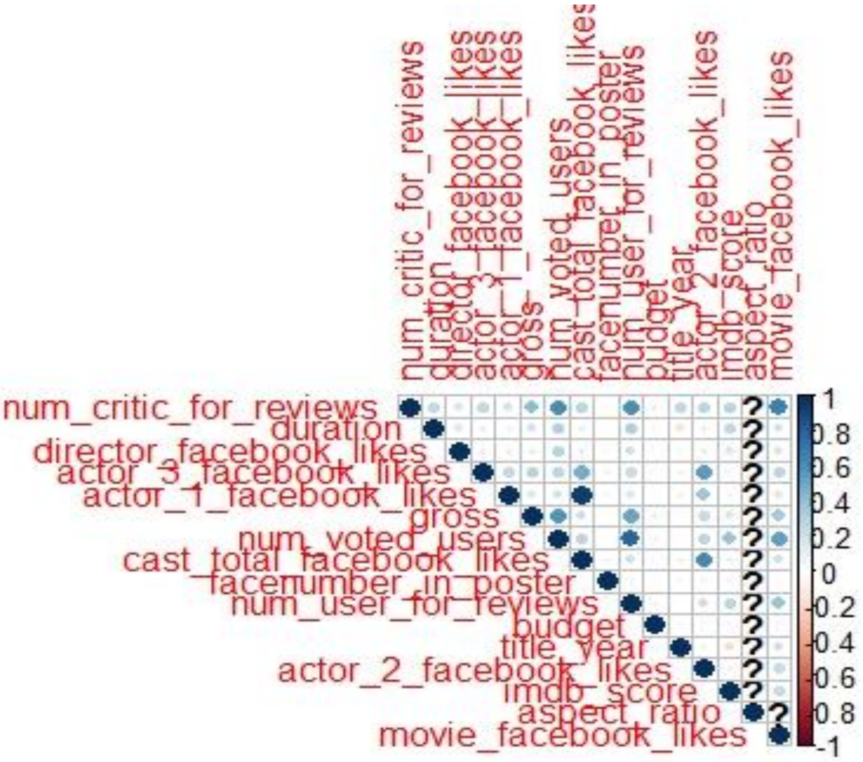


The above graph was done in R studio by using library(ggplot2) the command we used isggplot(data=movie_data,aes(x=gross,y=imdb_score,colour=factor(genres)))+stat_smooth(method=lm,fullrange = FALSE)+geom_point()From above plot, we can analyse that Action, crime movies are most likely to collect highest gross and had highest IMDB rating. Gross revenue is highly correlated with IMDB rating positively

Correlation plot for all variables in movie data set:

I have calculated correlation between all variables in Movie data and plotted a graph 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



The above graph was done in R studio.to do correlation we should intalllibrary(corrplot) I search in google for corrplot code I found. The code is corplot_values<-corrplot(corr_mat, method = "circle", type = "upper").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

Top ten movies:

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	Steam boy	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

The above table was done R studio**.** First I find outtop 10 movies with highest gross. I search in google I found The code is movie_gross<-head(movie_data[order(movie_data$gross, decreasing=TRUE),], 10). Second I calculated highest budget the code is budget_top<-head(movie_data[order(movie_data$budget, decreasing=TRUE),], 10).third I calculated lowest budget the code is budget_least<-head(movie_data[order(movie_data$budget, decreasing=FALSE),], 10) the data is mention in above table.

Top 10 director:

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

The above table was done R studio. First we find out top 10 director number of movies in movie data. I used a code isdirector<-data.frame(table(movie_data$director_name)). Second calculate highest gross revenue the code is directors_top<-head(director[order(director$Freq, decreasing=TRUE),], 11).

References:

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

Sensor data diary:

Week 4:

Day 1 and Day2:

We were given with the sensor data which assumption Column names, As column names were not mentioned in sensor data, we assumed first column in given data set are 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 cleansing:

Firstly, we have to analyse variables 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. I search in Google how to transposing data and I seen some videos in you tube .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.

Visualization:

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:

Firstly how to k-means clustering I search in Google I seen some data related to your project. The website is techtarget.*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.

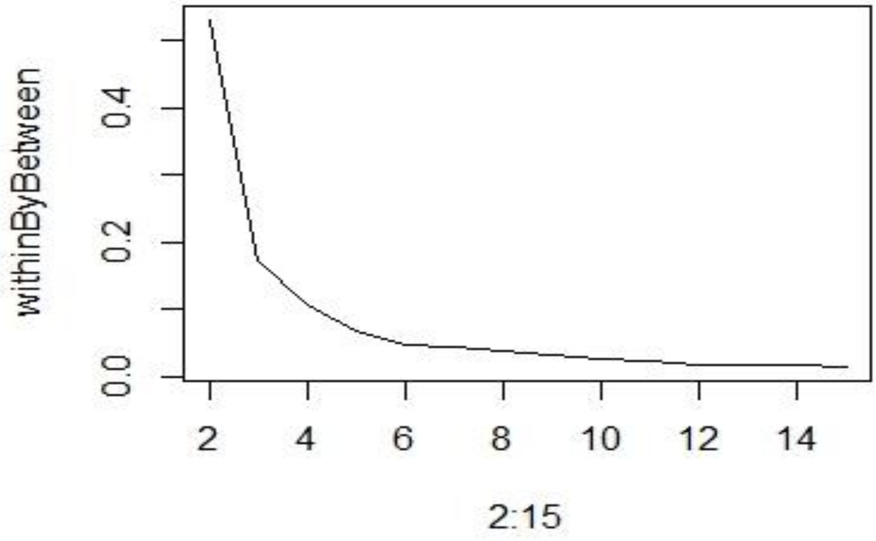
Each centroid of a cluster is a collection of feature values which define the resulting groups. Examining the centroid feature weights can be used to qualitatively interpret what kind of group each cluster represents.

Day 3 and Day 5:

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 the code is clust= kmeans(x=sensor,centers = 10), between SS and withSS values to find better explainability and plotted graph between and within SS to find ideal number of clusters to analyse data.



The visualization of above graph I used R studio. I search in google by know some data in website I used code is clust= kmeans(x=sensor,centers = 6) in R programming. 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

Reference:

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/