Project Report

First step into case studies

By

Shidharth Bammani

For

SRH Hochschule Heidelberg

23 November 2018

Abstract

The project report presents the data analysis for each of the four given datasets. The analysis of each dataset is explained in the report along with the chosen approach. The results are represented in a graphical way, the analysis if each data has been documented. The qualitative analysis approach to find the solutions to the analytical questions can be reflected through the report. The methodology is explained which describes the process of data analysis.

Contents

1.Funeral Dataset

The given dataset presents the findings of the qualitative analysis of given data set Funeral. The dataset contains the data of the people who lost their lives and the date of their funeral. The data can be cleaned and analysed by using Excel as tool.

Excel can be used to open the dataset present in text file and convert it into readable format. The variables are not mentioned in the dataset, so we have to guess the variables on the basis of data present. The dataset may also contain blank values and duplicates for which excel can be efficiently used to clean the data and analyse it to make use of vast information and get meaningful insights out of them.

As per the understanding drawn from the data value, below are the list of data variables along with short description.

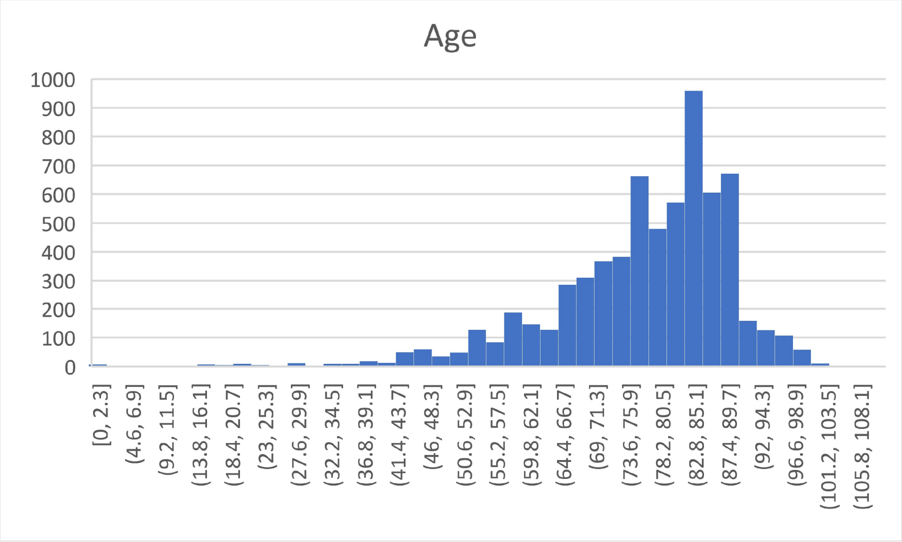
Address	The address of the deceased
Maidenname	Maiden name
DOB	Date of birth
DateofDeath	Date of death
FuneralDate	Date of funeral
Name	Firstname
Pincode	Pincode of the deceased
Street	Street name
Streetnumber	Street number
Surname	Last name
Time	Time

The overall 11,500 data rows are present after importing the text file, this data needs to be cleaned. The dataset can be cleaned by removing the blanks in the “date of birth” variable. After removing blank data rows, there were incorrect data mentioned in the “Date of birth” variable. So applying filter on that variable and removing all the incorrect year data. Also I the blanks present in the Date of death column, the date of funeral was assigned assuming that would not affect the data much. Now there are around 7500 data records left after performing the cleaning of data.

- Now the data is clean and analysis can be done. To analyse the data is to come up with potential questions that can be answered using the clean data. Below are the list of questions and the solutions to those question.

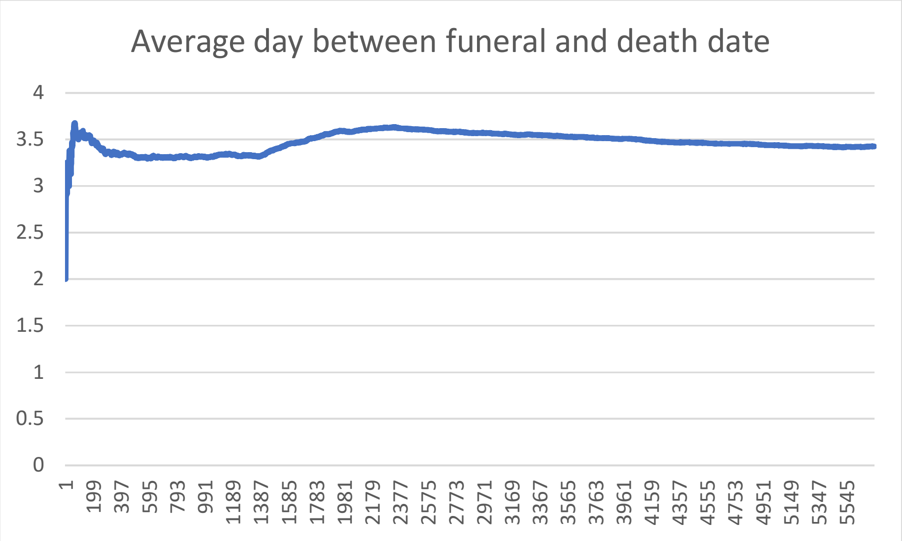
Q1. What is the average age of the people died?

- The average age of the people can be calculated by finding the age from the dataset, below graph describes number of people who lost their lives at a certain age group.



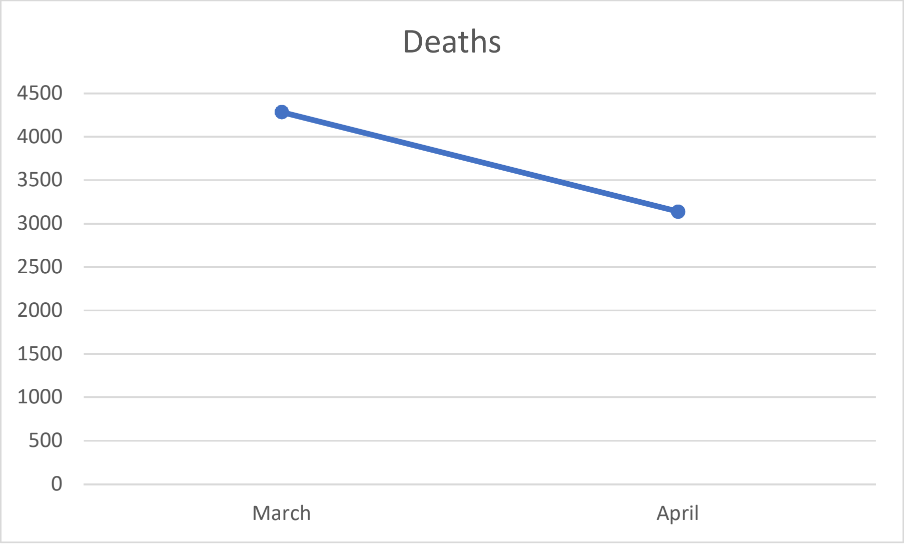
Q2. How long does it take for funerals to take place after the death of a person on an average?

- By further analysing the data we can get to know the average number of days between funeral and date of death. The below mentioned graph display that around 2 to 4 days on an average number of days after the death of a person.



Q3. In which month did most of the people lost their lives?

- The month in which majority of the people died can be calculated by taking count of no. of death happened in that particular month, below is the graph which explains that more number of people died in the month of march than April.



Additional variables will be more effective in analysis of the dataset, the additional variables are as follows "Sex"- by mentioning the gender, funeral date of entire year, Occupation, symmetry place of the dead to be buried, Cause of death.

By assuming we have all the above mentioned variables, below are the list of questions that could have been answered.

Q1. Which gender lost their lives more this year?

Q2. In which season did most of the people lost their lives?

Q3. What was the cause of death for most of the people this year?

Q4. What was the occupation of the people who died at a certain age group?

2.Titanic Dataset

This report presents the findings of the qualitative analysis of given data set Titanic. The data set contains the information of passengers who were present on the tragic mishap of Titanic ship. The qualitative analysis methods were used to clean the vast data and then extracting that data to get meaningful data. For cleaning and analysis, Excel was used as tool, since the given data is partially clean and the variables are clearly mentioned after we import the text file into excel. Excel makes it easier to plot charts, tables, also conditional formatting and filtering of variable columns can be done faster.

1. Data Cleaning:

Firstly, the given data set is a text file, so to make the data readable and to work on it, the text file was imported in to Excel. While importing the dataset the delimiters were set to tab as the data had too much space in between. After importing the file, the data is in readable format, the data quality is fine but it is still partially clean since some of the variables had blank values in them and some of the data value got shifted and there was mismatch of the variable and data value. The data columns were checked for duplicate values but found none.

Below is the list of variables found in the dataset along with the interpretation.

PassengerId	Unique passenger number
Survived	0 = no, 1 = yes
PClass	Passenger class (1, 2, 3)
Name	Family name, given name
Sex	male / female
Age	Alter
SibSp	Number of siblings and partner (spouse) aboard
ParCh	Number of parents and children aboard
Ticket	Ticket number
Fare	Ticket price
Cabin	Cabin number
Embarked	Entered in Southhampton (S), Cherbourg (C) or Queenstown (Q)

In total there were 950 data rows present, to clean this data, filtering out the variables serially looking for blank values. The 'PassengerId', 'Name', 'Sex', 'Ticket', 'Fare' and 'Embarked' variable had no missing data, so almost all the variables had data in them which indicates the quality of data is good. There were some blank values in the 'Survived', 'Pclass', 'Age' and 'Cabin' variable, but among these the 'Survived' and 'Pclass' variables are more useful.

By filtering the 'Survived' and 'Pclass' variables column there were 5, 2 data values missing respectively. To make the maximum use of given data instead of removing those records were searched by their names on the internet (URL: https://www.encyclopedia-titanica.org/) and the survival status data was obtained along with passenger class. The variable columns were checked for mismatches and were fixed.

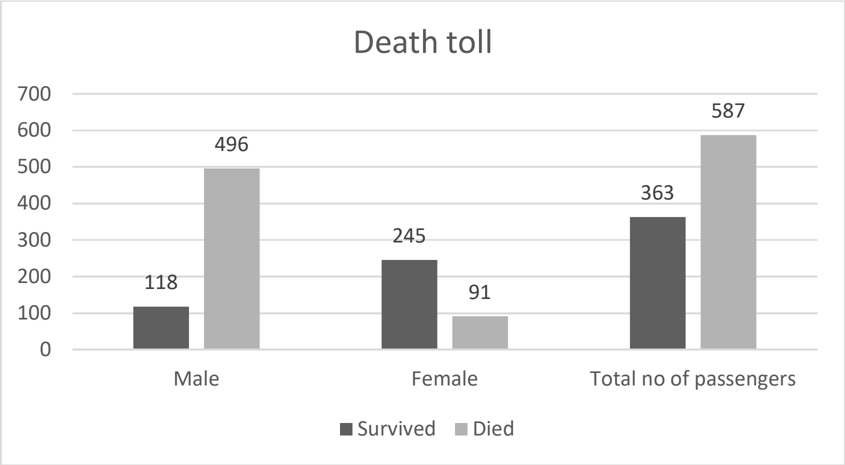
Now the data is clean and can start analysing the data and getting meaningful data out of them.

Data Analysis

To analyse the data is to come up with potential questions that can be answered using the clean data of the passenger list and get meaningful data out of them. Below are 5 questions listed along with the answers described in the form of graphical representations and short description of the process to answer those questions.

Q1. **What was the total no of the survivors on the ship and among them how many were male and female?**

The below graph explains the survivor and the death rate of the titanic disaster. The total death toll was 587 out of 950 passenger data in which we can see more number of females survived compared to male, this tells us that there were not enough lifeboats for passenger and crew.

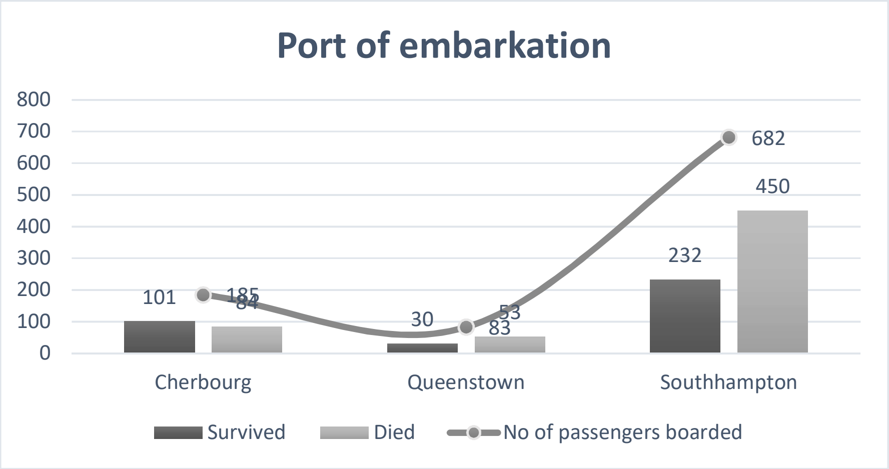


In the below table the survival status of the passenger. The table gives us the count of male, females and total number of passengers who lost their lives.

Status	Male	Female	Total no of passengers
Survived	118	245	363
Died	496	91	587

Q2. How many passengers embarked the ship from different ports and among them how many survived the disaster?

After analysing the data, the data can help us tell the number of passengers embarked the ship from which port and among them how many survived or lost their lives in the tragic accident. Below graph explains that a large number of passengers boarded the ship from Southampton port whereas the passengers embarked in Queenstown was the least. The data of the passenger names can further be of help to people finding their lost ones.

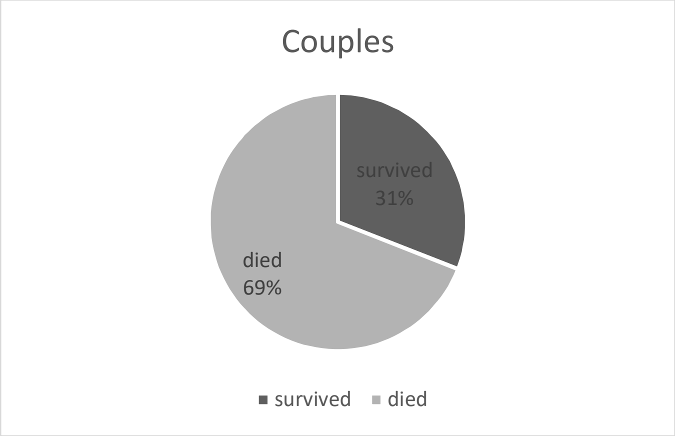


The below is the tabular data of the number of passengers boarding from different cities.

Port of embarkation	Survived	Died	No of passengers boarded
Cherbourg	101	84	185
Queenstown	30	53	83
Southampton	232	450	682

Q3. Among the survivors how many of them were couples and among which both of the partners survived or both of them lost their lives?

By filtering out the correct variable the data can tell us the number of couples in which both the partners survived the mishap, the number of couples in which both the partners died and the number of couple who sadly lost one of their partners. Below given is pie chart displaying the percentage of couple survived.



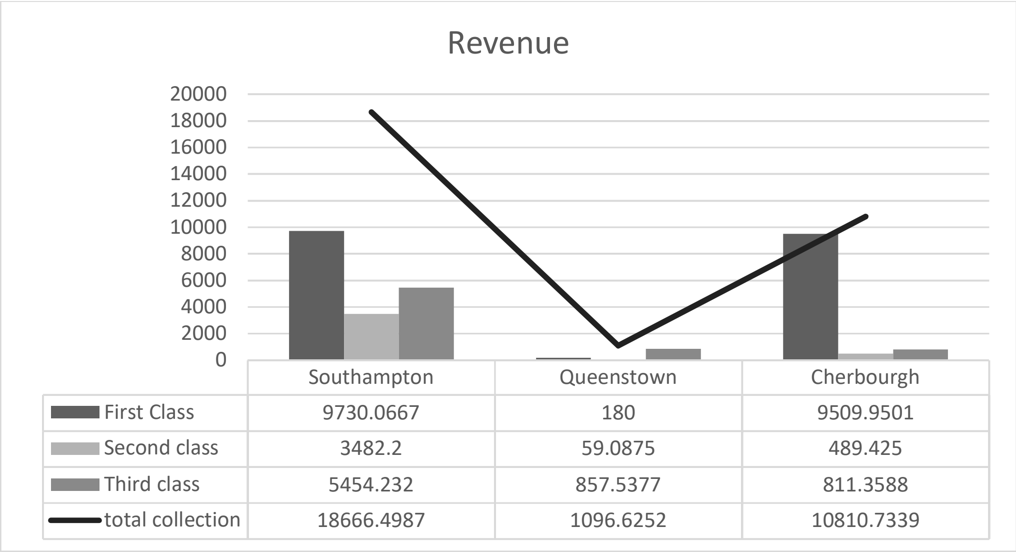
Below table provides the detailed number of passengers travelling with their partners and the ones who lost their loved ones in the accident.

Count	Description
108	Number of passenger who came along with their partners
54	Number of couples
13	Couples survived
29	Couples died

12	Couple who lost either of their partner
13	Male partner survived
41	Male partner died
39	Female partner survived
15	Female partner died

Q4. How much revenue was generated from each port?

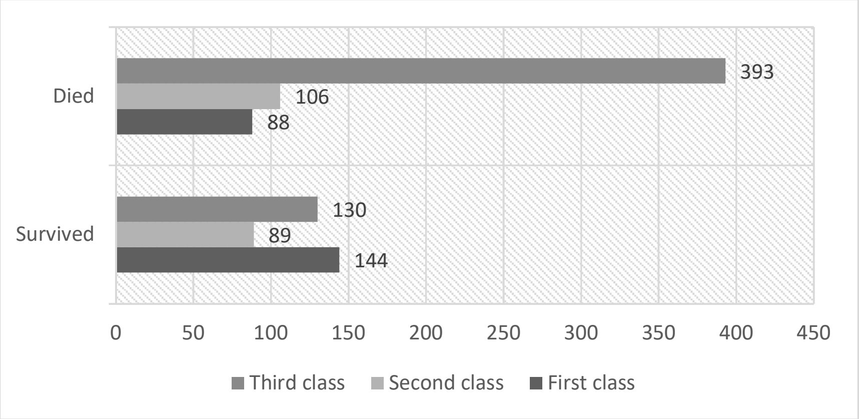
Since Titanic was the largest ship and most luxurious passenger ship of its time the revenue generated by the ship can be calculated by the fares of each passenger class and this figure can be plotted on the graph. I am assuming the currency as Euro in which tickets were sold since the ship sailed from Europe and was set off to New York. Below is the graph explain the prices of passenger class tickets in respect of the port of embarkation. Also there is table mentioning the exact amount collected from passenger respective of their class.



	Southampton	Queenstown	Cherbourgh
First Class	9730.0667	180	9509.9501
Second class	3482.2	59.0875	489.425
Third class	5454.232	857.5377	811.3588
total collection	18666.4987	1096.6252	10810.7339

Q5. Which class of passengers survived the most and which class lost most of their lives?

The below bar graph explains the passenger class who survived the most and the class which lost most of their passengers lives. This graph plot tells us that more maximum number of first and second class passengers survived whereas the most of the passenger travelling in third class lost their lives. This tells us that the preference to the lifeboats was given mostly to the upper class.



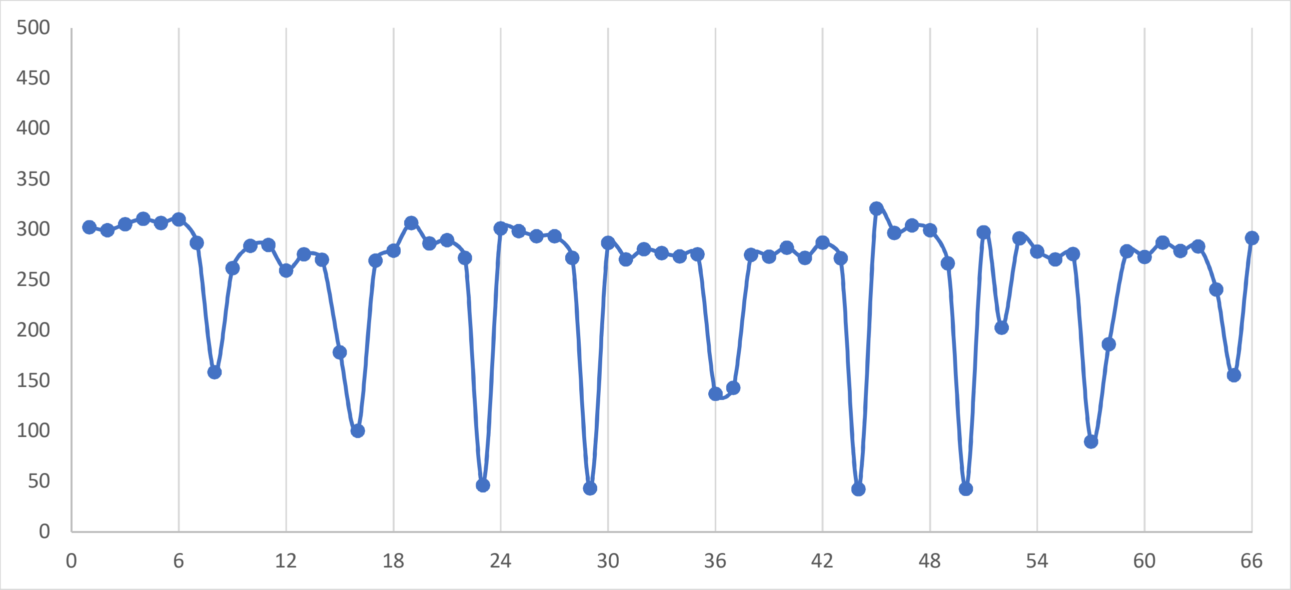
- I believe the data columns required to answer different analytical questions were mentioned but with respect to the family relation variables (i.e. SibSp and ParCh) some relations were not properly defined.

3.Sensor Dataset

The dataset given present data of sensor's readings, the variables are not given in the data set. The data file is in text format, to make the data readable the file is imported into excel. Using excel as tool to analyse the given sensor dataset.

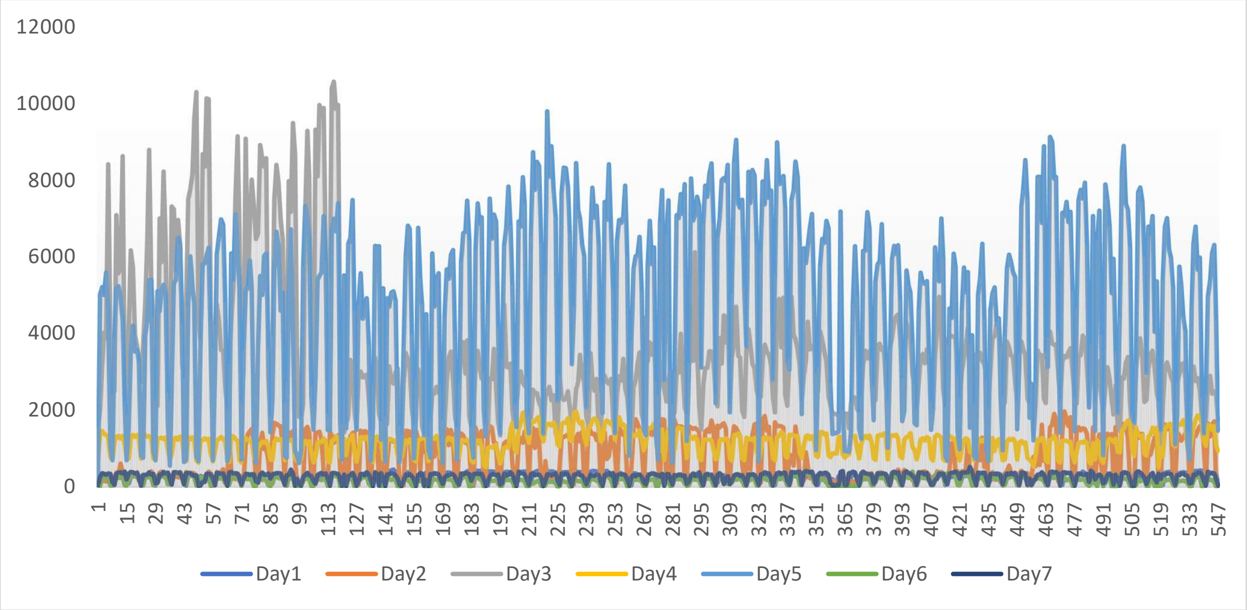
The variables are not mentioned in the data the values seem to be that of motion detector sensor of a corporate office.

In the below graph we have only plotted the first column to analyse the data, here we can see that after every five days the readings fall drastically i.e. for 6th and 7th day it drops and then again on 8th day it starts rising.



In the below graph the data values for seven column is used to plot the graph given below,

Here the the readings in day1 are higher assuming it is start of the week it keeps decreasing and finally on Day 7 the reading comes almost flat line. This shows pattern of a weekly working hours of any corporate office.


12000
10000
8000
6000
4000
2000
0
1
15
29
43
57
71
85
99
113
127
141
155
169
183
197
211
225
239
253
267
281
295
309
323
337
351
365
379
393
407
421
435
449
463
477
491
505
519
533
547
Day1
Day2
Day3
Day4
Day5
Day6
Day7

4.Movie Dataset

The given dataset presents the data of movies containing the names of the movies, its cast, director, review, budget etc. This dataset consists of large number of data values and variable columns. I have applied qualitative analysis method by cleaning the data and then analysing it to get meaningful output. The tool being used are Excel and SQL database, cleaning of data can easily done in excel and the conditional formatting of cell and filtering of data variables is faster, but to get solution for complex analytical question SQL is much easier to get solutions and saves a lot of time.

Below are the questions that can be answered while analysing the data.

1. *How to deal with missing data?*

- The dataset contains many variables which describe the data of movie based on names of movies, cast, director, revenue, budget and so on, all this data can be analysed to get meaningful insights. To analyse the data we need it to be clean in the sense the data values should be unique and should not have duplicates.
- In this dataset the column variable 'director_name' is one of the important variable based on which the analysis can be done to further extent. This variable had missing data and to deal with them, I filtered out the blank values and searched the names of movies on internet, I entered the data manually. The search result provided the names as well as came to know there were names of TV series mentioned in the movie data list.
- Analysing the data to find the answers to task questions, I applied filters to 'gross_budget', 'gross_revenue' variables looking for missing data, the data row with blank values were removed. This provided the clean data on which analysis can be done.

2. *How to correct wrong data or interpret data with unclear semantics?*

- While analysing the data I came across the data values which were incorrect. I found that the 'movie_title' had names of TV series in them, this data needs to removed. To search and remove all the data of tv series, I filtered the 'plot_keywords' by splitting the data column. I searched the keyword 'tvseries, sitcoms, series' and found the list of TV series and removed them.
- Two more column variables I found that contained data with unclear semantics were 'aspect_ratio', 'imdb_score'. For 'imdb_score' the data was provided as example: "7- Sept", I assumed that the month to be considered as the number of the month, and after formatting the data into number I found the imdb rating as '7.9'. Similar changes were made to 'aspect_ratio' variable column.

3. *How to pre-process the data?*

- By analysing the data, it indicates that some of the variable data columns needs to be processed and separated out in order to work on them to find the answers for the task question. The genre column had all the genre merged into a single column, this makes it difficult to find the relation between genre and other variable which can get interesting results. So the data column was split up in Excel by using 'Text to Column' operation, this splits the column into multiple columns. Similar changes were done to 'plot_keyword' variable data column.

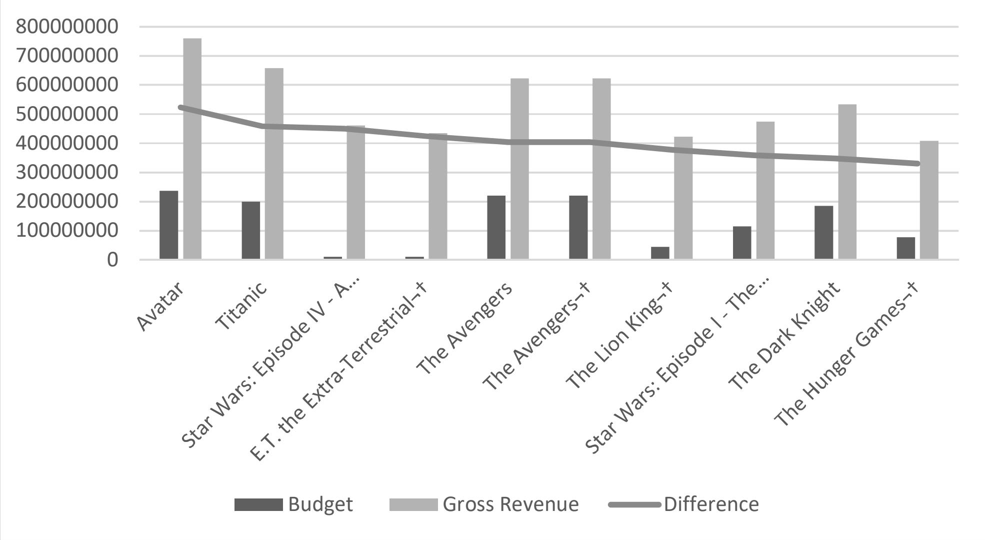
Below is the result of the analysis performed on the list of questions explained with the help of graphs and tables.

Q1. Top 10 movies with highest gross revenue, with largest budget, with smallest budget.

- Below is the table containing the top 10 movies with highest gross revenue with smallest budget.

Sr.no	Movie Titles	Budget	Gross Revenue
1	Paranormal Activity	15000	107917283
2	Tarnation	218	592014
3	The Blair Witch Project	60000	140530114
4	The Brothers McMullen	25000	10246600
5	The Texas Chain Saw Massacre	83532	30859000
6	The Texas Chain Saw Massacre	83532	30859000
7	El Mariachi	7000	2040920
8	The Gallows	100000	22757819
9	Super Size Me	65000	11529368
10	Halloween	300000	47000000

- Below graph displays the top 10 movies with highest gross revenue, with highest budget.



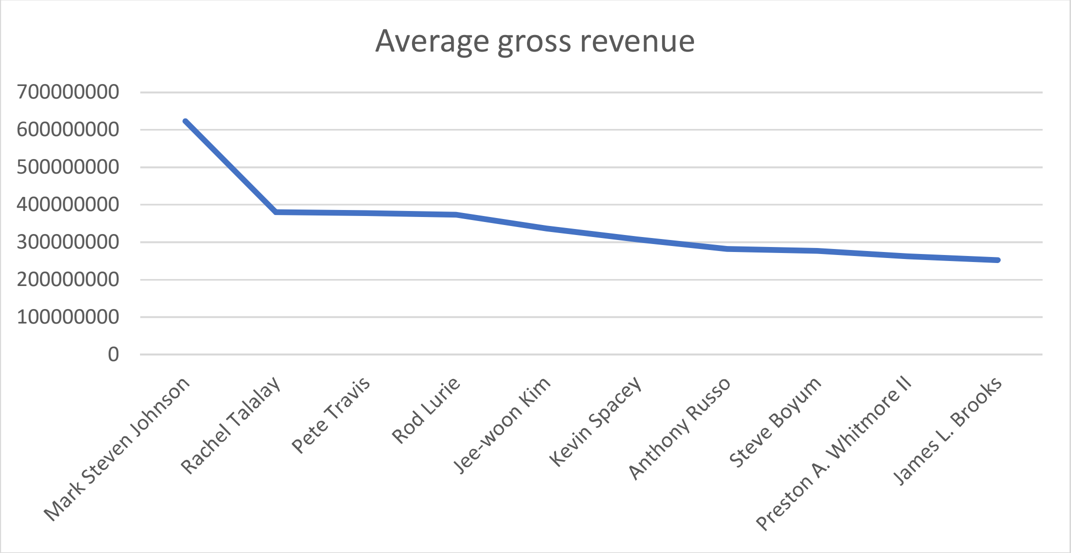
- Below is the table containing the top 10 movies with highest gross revenue, with highest budget.

Srno	Movie Title	Budget	Gross Revenue
1	Avatar	237000000	760505847
2	Titanic	200000000	658672302
3	Star Wars: Episode IV - A New Hope	11000000	460935665
4	E.T. the Extra-Terrestrial¬†	10500000	434949459
5	The Avengers	220000000	623279547
6	The Avengers¬†	220000000	623279547

7	The Lion King¬†	45000000	422783777
8	Star Wars: Episode I - The Phantom Menace	115000000	474544677
9	The Dark Knight	185000000	533316061
10	The Hunger Games¬†	78000000	407999255

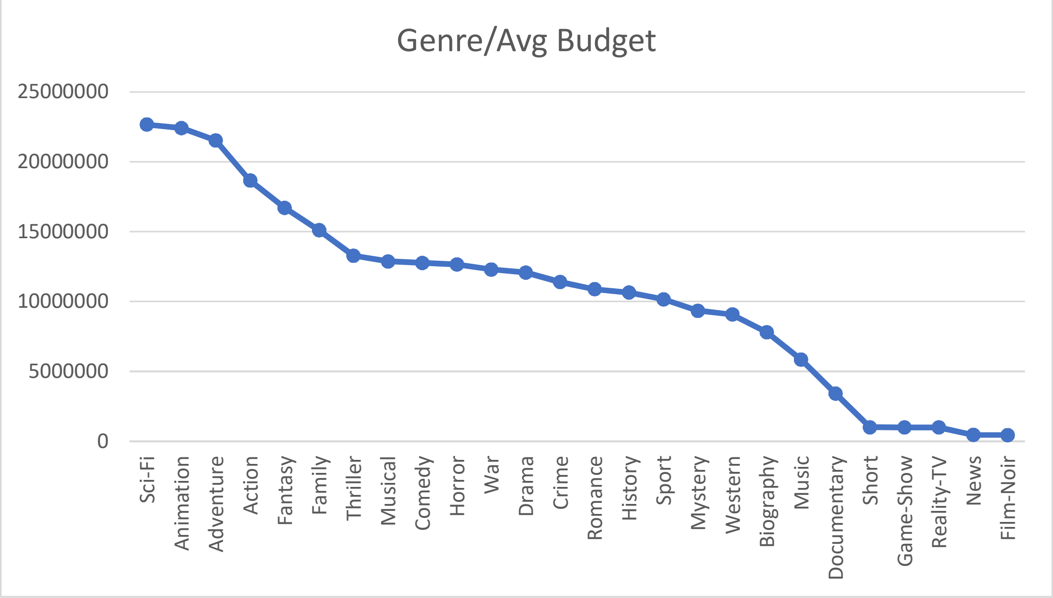
Q2. Top 10 directors according to number of movies, according to highest gross revenue.

- Below graph displays the names of top 10 directors with averages gross revenue of all their movies.



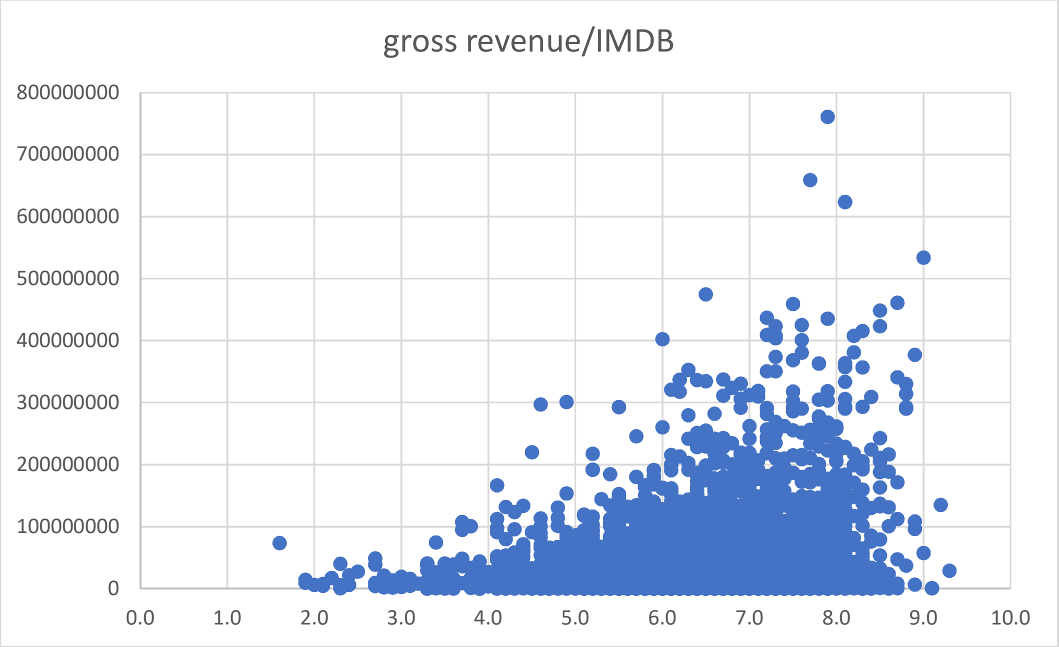
Q3. Relation between budget, gross revenue, IMDB rating and other variables.

- To find the relations between Gross budget and the genre of the film is very interesting. The below graph is plotted on the basis of genre of the film and the budget of the films. The analysis tells us that the genre of the film decide the type of budget required to produce the film.



Q4. Relation between gross revenue, facebook likes, number of reviews and other variables.

- The relation between the ‘gross revenue’ and ‘imdb’ variable interesting. We can see in the below graph that higher the IMDB rating the greater is the gross revenue of the film. Below the graph is plotted based on IMDB rating and the gross revenue of all the movies.



Bibliography

Dataset 2:

URL reference:

ticket fare: https://www.catawiki.com/catalog/miscellaneous/objects-items/vervoersbewijs/1192051-titanic-first-class-ticket-southampton-new-york-city

port data: https://www.thefreedictionary.com/port+of+embarkation

missing data: https://www.encyclopedia-titanica.org

Dataset 3:

Plotting single variable column: https://www.youtube.com/watch?v=7HpeFKUzPcI

Dataset 4:

IMDB: https://www.imdb.com/