Project Dairy

First step into case studies

By

Shidharth Bammani

For

SRH Hochschule Heidelberg

23 November 2018

1.Funeral dataset

The given dataset is provided in text file, to make the data readable we have to open the file in excel. The data file is opened in excel and after setting the delimiters the data is in readable format.

The variables are not mentioned in the dataset, by analysing the data value we can guess the variable of the data column. Below is the list of variables along with the interpretation.

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

- We have around 11,500 data records in the dataset, but this data needs to be cleaned since it may have duplicates and incorrect values in them. To check the data quality, we have to analyse the data values.
- After analysing the data, we come to conclusion that it needs lot of cleaning to be done. The variable 'Date of birth' had blanks in them, so we have to remove those data records. Some of the data records in 'date of birth' were incorrect, these data also need to be removed.
- The variable 'Date of death' have blank values in them, so assuming that the funeral date will be one day after the day of death, we fill out the data column based on date of funeral.
- The 'date of death' had time of funeral included in it, so we have to split the data column in to multiple column and format the column to date format.
- Now to filter out the data record which do not have names provided in them, we need to delete these records.
- After cleaning the data, we get around 7000 records on which we can work on and can get analytical questions and their solutions.

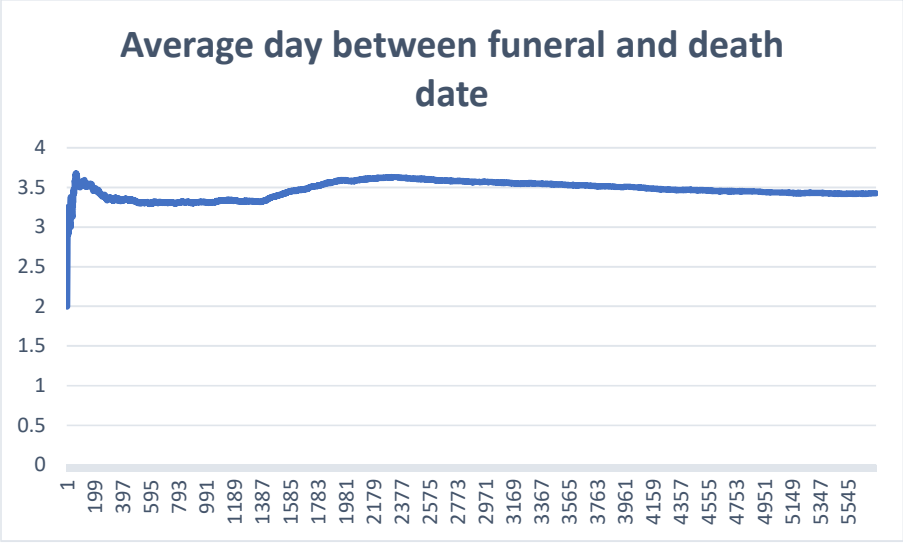
Next step is to get business logic question and its answers, below is the list of questions and the steps to obtain their answers.

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

- To find the average age first we need to find the age, we can make use of formula '=Datedif(DOB,DOD,"y")' to get the age of the people. After getting the age we can plot the age on to graph that will help us understand the average number of people who died at a certain age.

Next question is find the days between funeral and date of death.

- We can find the difference by using formula '=Datedif(DOB,DOD,"d")'.
- The below graph explains the average number of days taken between the date of death and funeral date.



2.Titanic dataset

First step is to export the text file into excel and convert it to CSV format. While importing the file the delimiters were set to Tab, as the data were separated by tab.

After The variables were clearly mentioned to interpret the data in separate columns.

- After importing the data, I found 950 data entries present. Though the overall data quality was good but it needed a little cleaning. Since some of the columns got shifted and some of the data entries were missing.
- To clean the data, I started by filtering out the 'Survived' variable column. I found around 5 data rows which did not mention whether the Passenger survived or not. So I google searched names of the 5 Passengers to check the survival status and edited the variable column for those data rows.

Passen	Survive	Pclass	Name	Sex
311	1	1	Hays, Miss. Margaret Bechstein	female
314		3	Hendekovic, Mr. Ignjac	male
709	1	1	Cleaver, Miss. Alice	female
919	0	3	Daher, Mr. Shedid	male
921	0	3	Samaan, Mr. Elias	male

url for 'Survived' variable info: https://www.encyclopedia-titanica.org

- Next I applied filters on all the variables and checked the blanks present in them, many of the variables such as sex, age, Sibsp, ParCh, Ticket, and Cabin had blanks entries in them but I ignored them as my business question did not focus on those data values.

I found two data row were blank for the variable 'Embarked', so I googled searched the names of those two entries and entered the data manually.

Now the data which I have is clean and can now further work on the analysis part.

Analysis:

I have listed out 5 business logic question based on the data and have started analysis of the data in a way that helps me get the answers for those questions. My first question is to calculate the overall survival and the death count of the passengers with respect to gender. I got the data by filtering the 'Sex' and the 'Survived' variable column. This data was used to plot the bar graph in the report. My second question is to count the survival status and the number of passengers boarding from different ports. This data was extracted by filtering out the 'Survived' and the 'Embarked' variable column. The third question is to extrapolate the number of couples in which both the partners either survived or died. I filtered out the

'Sibsp' variable by selecting '1' as this column has data of siblings or spouse, to further evaluate it I split the 'Name' variable into three columns as mentioned below.

Name	Salutation	Firstname
Arnold-Franchi	Mr	Josef
Arnold-Franchi	Mrs	Josef (Josefine Franchi)

I filtered out the 'Salutations' by selecting only 'Mr & Mrs' also I removed the siblings from the records, the name of the partners was same and even the ticket number were same, so this helped to get the count of passengers travelling with their partners.

My forth question is to calculate the total fare collected from the passengers based on the port they embarked. I filtered out the 'Embarked' and 'Pclass' variable and summed up the total revenue generated from the ticket fare. The final question I feel that can be answered was the passenger class which survived the most and the class which lost most of their lives. Similar to previous ways filtered the variable columns and plotted the data.

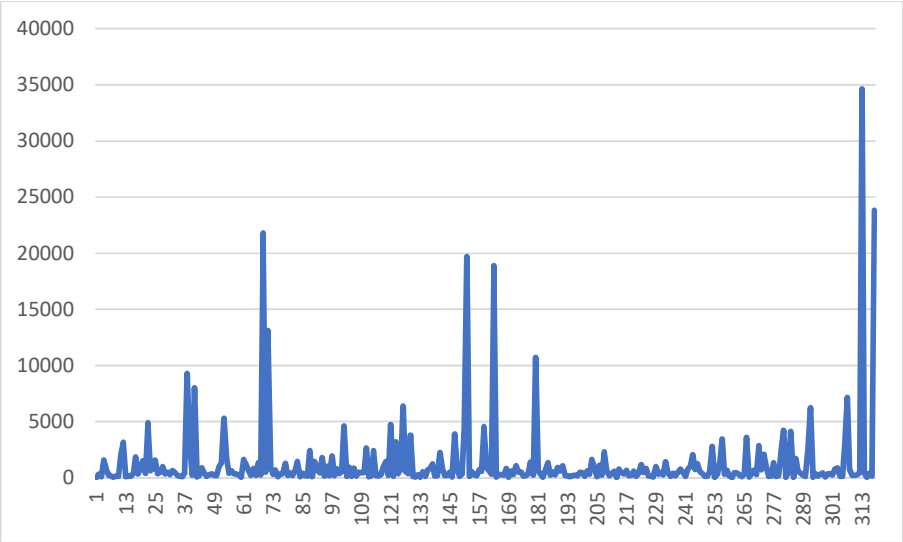
What all things went wrong while working on the Titanic dataset:

1. Earlier I had even checked the Space delimiter along with Tab but the column for Ticket got shifted since some of the data inside this column had space for eg: "PC 17599" and the other columns got affected. So I just used the Tab delimiter only.
2. Trying to find the relation between the Parent (either mother or Father) for a particular data record, but since the relations were not clearly defined, I stopped analysing the question further.

3.Sensor dataset

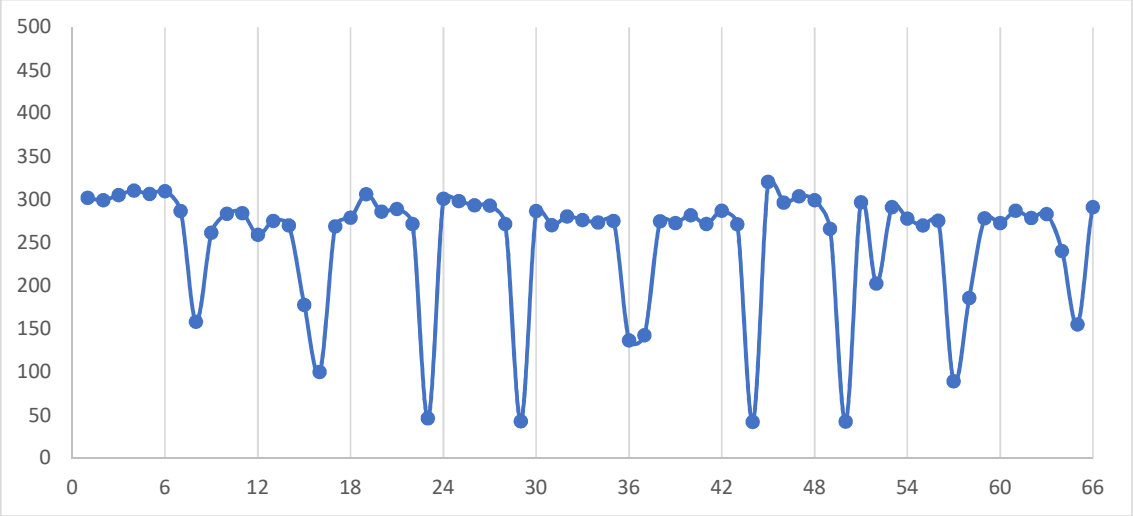
- The sensor dataset is given in text file first it needs to be imported in excel to make the data readable. After importing the data, we can see that the variables are not mentioned in the data columns.
- By analysing the data, we get to know that the data might be sensors readings.

- The data row had serial number specified, to find the relation or to get the understanding of the reading lets plot a graph for single column data.



By looking at the above graph the data column does not seem right, let's transpose the data table.
Let's plot a graph as per new changes just by using single variable column.

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.



The sensor might be a motion detector sensor installed in a corporate office.

4.Movie Dataset

- The movie data set is interesting to work on, the dataset is already present in excel file, so making use of excel to clean and analyse the data.
- Below is the list of variables provided in the data set along with the interpretation.

Color	cast_total_facebook_likes
director_name	actor_3_name
num_critic_for_reviews	facenumber_in_poster
Duration	plot_keywords
director_facebook_likes	movie_imdb_link
actor_3_facebook_likes	num_user_for_reviews
actor_2_name	language
actor_1_facebook_likes	country
Gross	content_rating
Genres	budget
actor_1_name	title_year
movie_title	actor_2_facebook_likes
num_voted_users	imdb_score
movie_facebook_likes	aspect_ratio

- Starting by cleaning the data, looking for duplicates and blank values in the main variables such as 'director_name', 'num_critic_for_reviews', 'gross', 'genre', 'movie title', 'plot_ keywords', 'budget'.
- Some of the variables such as 'plot_keyword' and 'genre' needs to be pre-processed in order to use them to find the relations between different variables data.
- Splitting the 'genre' variable by applying 'Text to columns' in excel, this operation splits the column into multiple columns providing separate list of genre for each movie. Similar operation will be carried out for the variable 'plot_keyword'
- Using excel as tool to remove duplicates, since the variable are clearly mentioned, filtering out the data and looking for blank values in the variable data column. The variable 'director' had many blank values, so I filtered them based on the movie names and searched them in internet, after this most of the director's names were found and manually added them in the data.
- While looking for director's names I also came to know that many of the 'movie_titles' variable had names of the tv series in them. So to remove those titles I filtered the earier processed column of 'plot_keyword' and searched for keywords such as 'tvseries', 'series', 'sitcom' and removed them.
- The data for 'aspect ratio' and 'imdb_rating' was unclear, for eg: '7-Sept' so I change the format of the column and made the data readable by assuming the month is

number according to the calender and got the value as '7.9' and changed the entire column data. Similar changes were made in 'aspect_ratio' column.

- Analysing the task question and getting results out of them:

Finding solution for question 1- 'To find the top 10 movies with highest gross revenue, with largest budget, with smallest budget'.

- I copied the data of four variables that are 'gross', 'movie_title', 'country' and 'budget' and pasted them in another sheet and removed the data rows with blanks. Now I calculated the factor of 'gross' and 'budget' and filtered them out in ascending order. The factor with least number was the movie with highest revenue and lowest budget. Below is the table.

gross	movie_title	country	budget	factor
107917283	Paranormal Activity¬†	USA	15000	0.000138995
592014	Tarnation¬†	USA	218	0.000368235
140530114	The Blair Witch Project¬†	USA	60000	0.000426955
10246600	The Brothers McMullen¬†	USA	25000	0.002439834
30859000	The Texas Chain Saw Massacre¬†	USA	83532	0.002706893
30859000	The Texas Chain Saw Massacre¬†	USA	83532	0.002706893
2040920	El Mariachi¬†	USA	7000	0.003429826
22757819	The Gallows¬†	USA	100000	0.004394094
11529368	Super Size Me¬†	USA	65000	0.005637777
47000000	Halloween¬†	USA	300000	0.006382979

- Next step is to find the top 10 movies with highest gross revenue and largest budget. The subtraction of gross revenue and budget gives me the profit margin of the film and arranging them in descending order gives me the list of movies with highest profit margin. Below is the list of movies with highest revenue, with highest budget.

gross	movie_title	country	budget	sub
760505847	Avatar	USA	237000000	523505847
658672302	Titanic	USA	200000000	458672302
460935665	Star Wars: Episode IV - A New Hope	USA	11000000	449935665
434949459	E.T. the Extra-Terrestrial¬†	USA	10500000	424449459
623279547	The Avengers	USA	220000000	403279547
623279547	The Avengers¬†	USA	220000000	403279547
422783777	The Lion King¬†	USA	45000000	377783777
474544677	Star Wars: Episode I - The Phantom	USA	115000000	359544677
533316061	The Dark Knight	USA	185000000	348316061
407999255	The Hunger Games¬†	USA	78000000	329999255

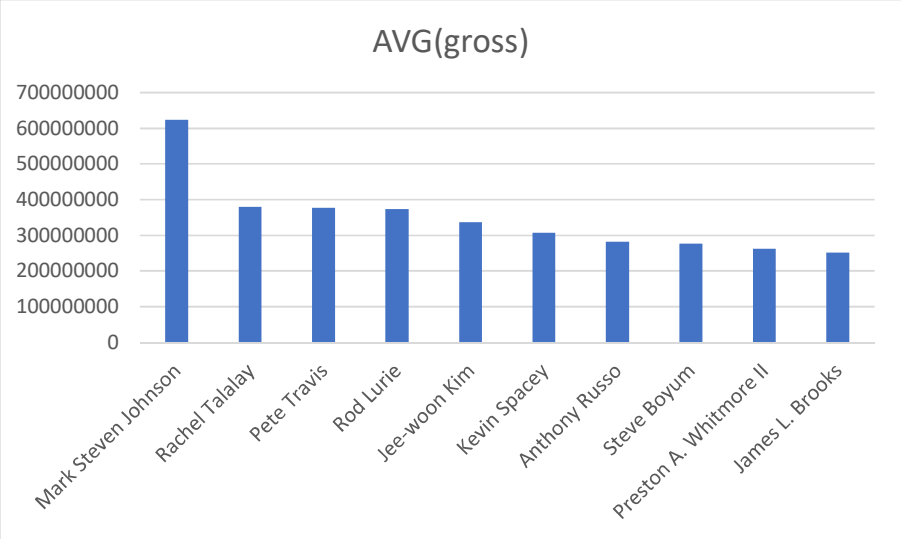
Steps to analyse the solution for question 2: 'Top 10 directors according to number of movies, according to highest gross revenue.'

- First step I copied the data of four variables that are 'gross', 'director', 'country' and pasted them in another sheet and removed the data rows with blanks. Now I applied conditional formatting on 'director' variable by highlighting the duplicate value then arranged them in ascending order. To get the count of duplicate value I applied formula "=COUNTIF(B:B,B2)" and got the no. of movies directed by same person. Below is the list of director's name along with number of movies directed by them.

Srno	director_name	Count
1	Steven Spielberg	25
2	Clint Eastwood	19
3	Woody Allen	19
4	Martin Scorsese	18
5	Ridley Scott	17
6	Steven Soderbergh	16
7	Tim Burton	16
8	Renny Harlin	15
9	Spike Lee	15
10	Barry Levinson	13

Step2: To find the top 10 directors according to highest gross revenue.

I used SQL database to calculate the average revenue of all the movies of a particular director. By using the previous sheet used in step 1, I uploaded the sheet in SQL database and wrote query to calculate the average of all the movies of the directors. After finding the average the maximum average were the directors with highest gross revenue. Below is the graph which displays the name of directors and average gross revenue of the movies they directed.



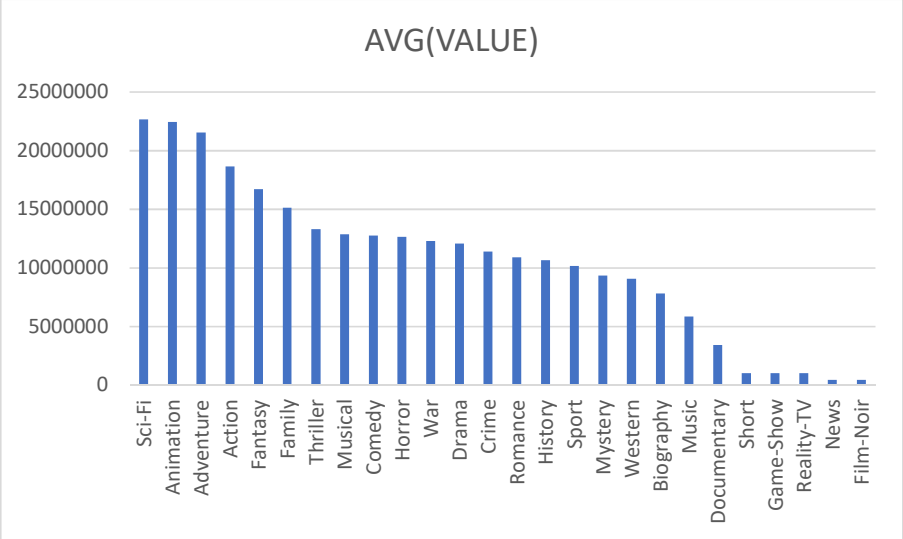
Steps to analyse the solution for question 3: 'Relation between budget, gross revenue, IMDB rating and other variables'

- I started to look for relation between the data variables which are interesting and through which meaningful results can be obtained. First I decided to find the relation between the 'gross revenue' and 'num_critic_for_review' . What I thought of is we can plot a graph through which we can come to know that more the number of critics for reviews means more the movie is being talked about and the number of people watching the movie will be greater hence the gross revenue of the movie will be highest. But the values to plot the graph were not enough to get proper presentation of the data.
- So I decided to find relation between genre and budget of the movie. The genre column had been processed in the earlier steps so I copied the columns of the movie_title, genre and the budget and pasted them in separate sheet.
- Next step is to divide the budget equally into the number of genres mentioned for a particular movie and adding the amount into each genre.
- Below is the table provided for understanding the steps.

genres	g1	g2	g3	g4	g5	budget	c1	c2	c3	c4	c5	c6	t	avg
Comedy	Drama	Horror	Sci-Fi			12215500000	1	1	1	1	0	0	4	3053875000
Crime	Drama					4200000000	1	1	0	0	0	0	2	2100000000
Drama	Romance	War				2500000000	1	1	1	0	0	0	3	833333333

- In total there are maximum six genre mentioned for a particular movie, here the genre has been split up and the c1, c2, c3, c4, c5, c6 are the count if the genre is present, the budget is divided as per the no of genre and the amount is assigned to the genres present in that record. The formula to calculate count is "IF(A2="",0,1)"

- The next step is to add up all the values of the genres by using formula “=IF(A3=A3,A3&" "&O3,0)” . After getting the total of all the averages of each genre, I used SQL db to get the average of the genre list.
- Below is the graph plotted based on the average value of each genre and the next is the table containing the values. This tells us that the budget of the movie depends on the genre of the film.



GENRE	AVG(VALUE)
Sci-Fi	22671397.3
Animation	22420627.73
Adventure	21544552.86
Action	18665962.13
Fantasy	16710746.07
Family	15109498.9
Thriller	13292220.77
Musical	12874545.88
Comedy	12767633.72
Horror	12668409.12
War	12301463.46
Drama	12077261.25
Crime	11404542.4
Romance	10885244.79
History	10652099.04
Sport	10162144.02
Mystery	9347636.039
Western	9078447.862
Biography	7820247.163
Music	5855561.543
Documentary	3416740.153
Short	1003250
Game-Show	1000000
Reality-TV	1000000
News	456500
Film-Noir	449552.1833

Finding solution for question 4- 'Relation between gross revenue and IMDB rating'.

- First step is to copy the gross revenue and imdb variable data into new sheet.
- Plot the graph based of revenue against IMDB rating.