CA682 Data management and visualisation

Name	Kashyap Krishnamurthy
Student Number	18210248
Programme	MSc in Computing (Data Analytics)
Module Code	CA682
Assignment Title	Data Visualisation
Submission date	17-December-2018
Module coordinator	Dr Suzanne Little

I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. I understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. I have read and understood the Assignment Regulations set out in the module documentation. I have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the source cited are identified in the assignment references. This assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study.

I have read and understood the referencing guidelines found recommended in the assignment guidelines.

Name: Kashyap Krishnamurthy Date: 17-December-2018

Table of Contents

1. **INTRODUCTION**

2. **DATASET**

   2.1 Data Source
   2.2 Accessing the data
   2.3 Data Understanding

3. **SOLUTION OUTLINE**

4. **DEVELOPMENT**

   4.1 Data Preparation
   4.2 Visualisation
   4.3 Webpage
   4.4 Execution

5. **ANALYSIS**

   5.1 Insights
   5.2 Principles implemented
   5.3 Scope of improvement

6. **GITLAB**

1. Introduction

Towering at altitudes of over 8500 meters, the Mighty-Himalayas, are rightly called the roof of the world. The third largest deposit of ice and snow in the world, after Antarctica and the Arctic, are in the Himalayan range. Mount Everest at 8848 meters is not only the highest peak in the Himalayas, but the highest peak on the entire planet. Himalayan mountaineering expeditions began in the 1880s, in the early 20th century the number of mountaineering expeditions increased markedly. Easier access to the mountains brought increasingly large numbers of climbers and hikers into the region. By the late 20th and early 21st centuries, the annual number of mountaineering expeditions and tourist excursions to the Himalayas increased remarkably.

Leveraging the data from The Himalayan Database, this report attempts to investigate if Himalayan Expeditions have gotten safer in the last century of climbing. Major expeditions are aimed for peaks at altitudes between 5400 to 8848 meters, and a common assumption is that expeditions to the highest peaks claim the most number of lives. The second aim of this report is to use a data driven approach to validate this theory. A third visualisation is included to highlight the mountains that see the highest number of climbers in the region.

Being a mountaineer myself, and having climbed in the Himalayas on numerous occasions, the results from this analysis will also help me gain a better understanding on the subject, at a personal level.

2. Dataset

This section details the dataset used for this analysis.

2.1 Data Source

The Himalayan Database is a compilation of records for all expeditions that have climbed in the Nepal Himalaya. The database is updated bi-annually and is published by The Himalayan Database. It is available for download from this website. The data are based on the expedition archives of Elizabeth Hawley, a journalist in Kathmandu, and it is supplemented by information gathered from books, alpine journals, magazines, and correspondence with Himalayan climbers. The data covers all expeditions from 1905 through 2017 to more than 450 significant peaks in Nepal. Each expedition record contains detailed information including dates, routes, camps, use of supplemental oxygen, successes, deaths and accidents. The volume of data that has been collected is comprehensive. The data record summary is as follows:

- Peaks records: 466
- Expedition records: 9797
- Member records : 71,000+
- Literature records: 14,077

2.2 Accessing the data

The Himalayan Database is a Microsoft Visual Foxpro 9 program with the required tables embedded within the application. When attempting to access the data, two challenges were faced -

1. The Microsoft Visual Foxpro 9 program is a .exe application, my machine running a Unix based OS, wouldn't run the application.
2. The design of the query within the database was such that, expedition records of only one peak can be exported for a single request, with over 450 peaks, this was not a viable option.

To counter these challenges, the backend files for each table - .CDX, .DBF and .FPT - were extracted, processed and converted to .CSV format using python.

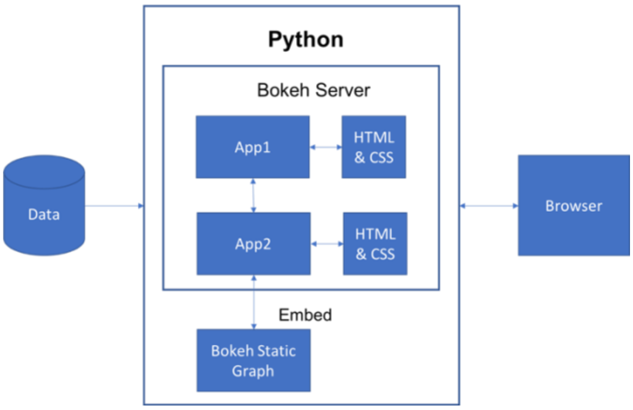
2.3 Data Understanding

To perform the three areas of analysis, two tables within The Himalayan Database, ‘exped’ and ‘peaks’, have been used.

- The exped table contains extensive information on the expeditions. The table contains 67 columns with information ranging from Date, Success of expedition, Ropes used, Oxygen, Total Climbers, Climber Deaths to name a few.
- The peaks table contains 22 columns with information related to a given mountain. Information such as name, altitude, open for climbing or not and region are included.

3. Solution outline

Python is the chosen development language. The development has been carried out using the Spyder IDE. To achieve the two interactive visualisations, two bokeh server applications are developed. The static graph embedded within the bokeh application will be generated using a standalone python script that leverages bokeh libraries. HTML and CSS are used to present the application outputs via a web browser.



4. Development

This section details the various aspects of the visualisation development.

4.1 Data Preparation

Data is prepared in the following steps –

1. The required data, identified in two tables, are imported as dataframes and first level of column filtering is performed. The 89 columns available are reduced to 15 columns relevant for the scope of this analysis.
2. The two dataframes created from two tables are merged using the PEAKID as the foreign key.
3. The datatypes for the columns are defined to perform further mathematical processing.
4. App1 is to visualise the expeditions for the last century and hence the data is grouped by the year column. App2 which is to visualising how dangerous the highest mountains are, the data is prepared by grouping the dataframe by the altitude. Data for the static graph is prepared by aggregating the records with respect to the peakid.
5. Multiple columns that define the success of an expedition are aggregated and the percentage of success is calculated.
6. Multiple columns defining the number and type of climbers are suitably aggregated.
7. Columns with the fatality numbers are processed to identify fatality figures for different types of climbers. The data is also calculated in terms of percentage.
8. Except the altitude and year columns, all the original columns are dropped and 11 derived columns are retained.

9. The final dataframe is analysed to identify values that are incorrect. Example – an expedition climbed a few hundred meters about the summit of Mt.Everest and another expedition had a success rate of 400%. Such data points, caused by incorrect source data, are cleaned.

4.2 Visualisation

The scatter plot created by App1 and App2 share the same design, the X-axis is App1 has year data while the X-axis on App2 has the altitude data. The data displayed in Y-axis can be chosen from the 11 listed options, the chosen data points populate accordingly. The 11 options are also available to be visualised using a colour and a size scale. Widgets to pan, zoom, hover and reset the visualisation are also included. The data aggregated for the 6 most-climbed mountains are used to generate the nested bar graph visualisation.

Finally, Non-functional aspects such as themes and layouts are added to improve the aesthetics of the visualisation.

4.3 Webpage

The main-python code is integrated with HTML and CSS files such that the visualisation is embedded in a descriptive webpage. Webpage1, powered by App1, hyperlinks to webpage 2 that is powered by App2. The static graph is embedded within webpage2, however, the bokeh application loads the graph and not HTML.

4.4 Execution

The bokeh servers are initialised from the localhost on port 5006 and on port 8080 simultaneously. This allows both visualisations to be live at any given point and allows 'back' and 'forward' actions on the webpages.

5. Analysis

5.1 Insights

It is observed that though the number of records in the dataset are high, when aggregated, they form small figures. This is due to the nature of the activity - there are a limited number of people climbing in the Himalayas. There is also a large variance in the data, some expeditions have been a disaster while others fared average, and a few have gone through well. The two questions that are being analysed -

1. Have Himalayan Expeditions gotten safer in the last century of climbing?

When zoomed in to the data-points between 1940 and 2017, there is up-to 7% increase in the 'AllClimberDeathPercent' from 1940 to 1970. However, post 1970, there is steady fall in the percentage. The changes post 1970 can also be seen in the 'ExpeditionSuccessPercent', which rises from below 20% to over 70%. Looking beyond percentages, the 'TotalClimberDeaths' is now higher than ever, prior to 1970 there were under 7 recorded deaths but by 1980 the numbers were close to 35, a fivefold increase. This number continues to stay in the mid 30s up until 2015. It can also be observed that the rise in the number of deaths is more for the 'member' climbers as opposed to the more experienced 'hired'(Sherpas) climbers, this is a reflection of how inexperienced climber attempt Himalayan expeditions for glory – also known as 'summit fever'. However, when the 'TotalExpeditions' and 'TotalClimbers' are overlaid using colour and size, a clear trend can be observed. On selecting the facets on the Y-axis we can observe a sharp increase - from about 25 in 1970 to over 400 in 2015 for expeditions and around 250 climbers in 1970 to over 4000 in 2015. Post 2015 there is a general trend of a fall in the numbers across the data, which is probably due to the Nepal earthquake in 2015, that has affected subsequent expeditions. Hence it can be concluded that, though the number of people losing their lives have increased, Himalayan Expeditions have statistically gotten safer. This perhaps is because we understand the mountains better, climbing techniques have been improved, we have designed and developed better equipment.

(Between minutes 1:09 and 3:12 in the video, the above mentioned visualisation has been presented)

2. How deadly are the Eight-Thousanders?

Eight-Thousanders are the mountains in the Himalayas that have summits above 8000 meters in altitude. It is assumed that expeditions to the eight-thousanders cause the maximum number of fatalities. When the 'TotalClimberDeaths' is visualised on the Y-axis, it can be seen that summits above 8000 meters are indeed causing the highest number of deaths. At close to 300 recorded fatalities, Mt. Everest is the highest and 4 other 8000 meter plus summits causing 50-80 fatalities. Visualising the 'ExpeditionSuccessPercent' also reveals that 8000 plus meter expeditions have about 60% success at best. When the visualisation is overlaid with 'TotalClimbers' and 'TotalExpeditions' data using the colour and size scale, it can be observed that the eight-thousanders have the highest number of expeditions and climbers. To analyse further, the 'TotalClimbers' data is visualised on the Y-axis. It can be observed that the eight-thousanders, though few in number, see a large share of the climbers, especially Mt. Everest, with a staggering 34,000+ climbers. Having the 'TotalExpeditions" on the Y-axis also reveals that 8 of the top 10 mountains with the highest number of expeditions are eight-thousanders, Mt. Everest being the highest yet again with close to 2000 recorded expeditions. Visualising the 'AllClimberDeathPercent' gives the right picture, on zooming into the range between 6500 meters and 8900 meters, it can be observed that there is in fact a steady fall in the percent of death right after the 8000 meter mark. Summits between 6500 meters and 7500 meters have a higher kill percent, 7000 being the highest at over 12%. Hence, it can be concluded that, eight-thousanders do kill the highest number of mountaineers, however, the figure is due to the high number of summit attempts being made on these mountains. It is in fact summits between the 6500 and 7500 meters that are more dangerous.
(Between minutes 3:20 and 4:52 in the video, the above mentioned visualisations has been presented)

3. The mountains that are climbed highest in the region?
The nested bar graph reveals that for the years between 2015 and 2017, though expeditions were affected due to the Nepal Earthquake, there were still a considerable number of summit attempts. It can be observed that Mt. Ama Dablam, Mt. Everset and Mt. Manasalu are the peaks with the highest climbers in the region. It is interesting to note that the highest mountain, Mt. Everest, sees the highest number of climbers.

5.2 Principles implemented

The principles implemented for the interactive visualisations are detailed below –

1. **Choice of Visualisation:** Visualisation being carried out is of the relational type and in accordance to Andy Kirks CHRTS methodology, the scatter plot has been chosen for the two interactive plots.
2. **Colour Scheme:** The following principles have been implemented in relation to the colours used –
   - Colour for scatter plot data points – colour is best used for categorical data and though all the data types in this visualisation are continuous, the colour option has still been employed to gain a surface level pattern of high and low. If there is a pattern observed, the same can be viewed on the Y-axis for a more detailed analysis.
   - Hue colours – hue colours have been chosen as the colour range since the saturation and the lightness scales were harder on the eyes to observe. The hue is also an internationally accepted pattern.
   - Colour Blindness / Black & White – The basic theme of the visualisation is considers the colour blindness and the black & white functionality factors.. By virtue of being an interactive graph, the colour option can be completely omitted should there be a need and the size scale can be used in place.
   - Contrast – the contrast between the webpage background, the graph background and the data points have been considered and factored in accordingly. Strong colours have also been used for small areas.

3. Other implemented principles –
   - Title on the top-left
   - Labels used appropriately and sparingly
   - Accurate proportions
   - Axis intervals are evident

- Guidelines are grey
- The visualisation has a boundary
- No decoration present

The principles implemented for the nested bar chart are –

1. Choice of Visualisation: Visualisation being carried out is of the categorical type and in accordance to Andy Kirks CHRTS methodology, the bar plot has been chosen for the two interactive plots.
2. Colour Scheme: The bars are coloured and are similar for a given year, however, the visualisation can be read by the colour blind and used in black & white. The contrast between the webpage background, the graph background and the data points have been considered and factored in accordingly. Strong colours have also been used for small areas.
3. Other principles –
   - Title on the top-left
   - Labels used appropriately and sparingly
   - Accurate proportions
   - Axis intervals are evident
   - Guidelines are grey
   - The visualisation has a boundary
   - No decoration present

5.3 Scope of improvement

Due to constraints such as technical limitations and time, the current version of the visualisation solution doesn't include the following functionality.

1. Varying X-axis – It can be observed that the X-axis of the visualisation has a drop down function, the initial plan was to have a varying X-axis for year and altitude. However, passing two different dataframes within the visualisation function needed additional development efforts for which the time available was a challenge. Hence, the work around chosen was to have two bokeh servers running applications parallelly, linked via the webpages. However, the drop-down menu for the X-axis has been retained, such that the required development can be carried out when plausible.

2. App2 - the nested bokeh bar graph that is loaded by App2 is an image that has been exported by a third python script to the 'static' directory of the bokeh server2. The image resource is then imported and loaded by App2. The initial plan was to have the data for static graph and the graph itself processed within App2, however, by doing so, the interactive element of the primary scatter plot was affected and I have not yet been able identify a solution, hence the above mentioned workaround.

3. Live data import – The Himalayan Database is updated bi-annually, and hence the visualisation can continue to be relevant if the data source is not static. The initial plan was to explore the possibilities of an Api to import the data, however, the website does not have any Api endpoints.
   The work around thought of was to scrape the website, download the data to a designated directory and have the python scripts pointed to this directory, however, the data available on the website is a Microsoft Visual Foxpro 9 program, this brought about unforeseen complexities for scrapping. Even accessing the standard downloaded file of the database had its challenges and the same has been briefed in the Section 2.2 of this report.

6. Gitlab

The visualisation source code, webpage code and the data can be found in the following repository.