POWER BI PROJECT

Requirements

- Select one dataset
- Clean, transform and structure the dataset
- Create a report using Power BI desktop
- Include at least five visuals in your Power BI report
- Create a dashboard using Power BI Service
- Build a write-up (Word document/Google docs) stating your questions (Problems) and explaining your findings (Insights and Analysis) from the Power BI report and dashboard

POWER BI Datasets

1. Chocolate Bar Ratings

Context

Chocolate is one of the most popular candies in the world. Each year, residents of the United States collectively eat more than 2.8 billion pounds. However, not all chocolate bars are created equal! This dataset contains expert ratings of over 1,700 individual chocolate bars, along with information on their regional origin, percentage of cocoa, the variety of chocolate bean used and where the beans were grown.

Flavours of Cacao Rating System:

- 5= Elite (Transcending beyond the ordinary limits)
- 4= Premium (Superior flavour development, character, and style)
- 3= Satisfactory (3.0) to praiseworthy (3.75) (well made with special qualities) •
  2= Disappointing (Passable but contains at least one significant flaw)
- 1= Unpleasant (mostly unpalatable)

Each chocolate is evaluated from a combination of both objective qualities and subjective interpretation. A rating here only represents an experience with one bar from one batch. Batch numbers, vintages and review dates are included in the database when known. The database is narrowly focused on plain dark chocolate with an aim of appreciating the flavours of the cacao when made into chocolate. The ratings do not reflect health benefits, social missions, or organic status.

Flavour is the most important component of the flavours of Cacao ratings. Diversity, balance, intensity, and purity of flavours are all considered. It is possible for a straightforward single note chocolate to rate as high as a complex flavour profile that changes throughout. Genetics, terroir, post-harvest techniques, processing and storage can all be discussed when considering the flavour component.

Texture has a great impact on the overall experience, and it is also possible for texture related issues to impact flavour. It is a good way to evaluate the makers vision, attention to detail and level of proficiency.
Aftermelt is the experience after the chocolate has melted. Higher quality chocolate will linger and be long lasting and enjoyable. Since the aftermelt is the last impression you get from the chocolate, it receives equal importance in the overall rating.
Overall Opinion is really where the ratings reflect a subjective opinion. Ideally it is my evaluation of whether or not the components above worked together and an opinion on the flavour development, character and style. It is also here where each chocolate can usually be summarized by the most prominent impressions that you would remember about each chocolate.

Acknowledgements

These ratings were compiled by Brady Brelinski, Founding Member of the Manhattan Chocolate Society.

2. Life Expectancy (WHO)

Context

Although there have been lot of studies undertaken in the past on factors affecting life expectancy considering demographic variables, income composition and mortality rates. It was found that effect of immunization and human development index was not considered in the past. Also, some of the past research was done considering multiple linear regression based on data set of one year for all the countries. Hence, this gives motivation to resolve both the factors stated previously by formulating a regression model based on mixed effects model and multiple linear regression while considering data from a period of 2000 to 2015 for all the countries. Important immunization like Hepatitis B, Polio and Diphtheria will also be considered. In a nutshell, this study will focus on immunization factors, mortality factors, economic factors, social factors and other health related factors as well. Since the observations this dataset are based on different countries, it will be easier for a country to determine the predicting factor which is contributing to lower value of life expectancy. This will help in suggesting a country which area should be given importance in order to efficiently improve the life expectancy of its population.

Content

The project relies on accuracy of data. The Global Health Observatory (GHO) data repository under World Health Organization (WHO) keeps track of the health status as well as many other related factors for all countries The datasets are made available to public for the purpose of health data analysis. The dataset related to life expectancy, health factors for 193 countries has been collected from the same WHO data repository website and its corresponding economic data was collected from United Nation website. Among all categories of health-related factors only those critical factors were chosen which are more representative. It has been observed that in the past 15 years, there has been a huge development in health sector resulting in improvement of human mortality rates especially in the developing nations in comparison to the past 30 years. Therefore, in this project we have considered data from year 2000-2015 for 193 countries for further analysis. The individual data files have been merged into a single dataset. On initial visual inspection of the data showed some missing values. As the datasets were from WHO, we found no evident errors. Missing data was handled in R software by using Missmap command. The result indicated that most of the missing data was for population, Hepatitis B and GDP. The missing data were from less known countries like Vanuatu, Tonga, Togo, Cabo Verde etc. Finding all data for these countries was difficult and hence, it was decided that we exclude these countries from the final model dataset. The final merged file (final dataset) consists of 22 Columns and 2938 rows which meant

20 predicting variables. All predicting variables was then divided into several broad categories: Immunization related factors, Mortality factors, Economical factors and social factors.

Acknowledgements

The data was collected from WHO and United Nations website with the help of Deeksha Russell and Duan Wang.

3. US Police Shootings

Context

In the recent US police killings, a hot topic came into being, "Racism".
This data has been gathered to take out some insights and analyse the story around racism in America.

Content

It contains basic data about people like their name, age, gender, and race. Along with it, is the shooting/killing information, like date of event, where it happened? how they were shot? did they attack? Were they holding weapons? Did they show any mental illness? Was the policeman wearing a camera/was the incident recorded? Did the suspect flee? Apart from that, a category column holds type of weapon used by the suspect

4. Udemy Courses

About this Data

This dataset contains **3.682 records** of courses from **4 subjects (Business Finance, Graphic Design, Musical Instruments and Web Design)** taken from Udemy.
Udemy is a massive online open course (MOOC) platform that offers both free and paid courses. Anybody can create a course, a business model by which allowed Udemy to have hundreds of thousands of courses.

This data has the following column names (explanation attached).

course_id - Course ID
course_title - Course Title
url - Course URL
is_paid - Whether the course is free or paid
price - Course Price
num_subscribers - Number of subscribers
num_reviews - Number of reviews
num_lectures - Number of lectures
level - Course difficulty
content_duration - Duration of all course materials
published_timestamp - Date that the course was published.
Subject- Course subject

5. Video Game Sales

Analyse sales data from more than 16,500 games. This dataset contains a list of video games with sales greater than 100,000 copies. Fields include

- Rank - Ranking of overall sales
- Name - The games name
- Platform - Platform of the games release (PC, PS4, etc.)
- Year - Year of the game's release
- Genre - Genre of the game
- Publisher - Publisher of the game
- NA_Sales - Sales in North America (in millions)
- EU_Sales - Sales in Europe (in millions)
- JP_Sales - Sales in Japan (in millions)
- Other_Sales - Sales in the rest of the world (in millions)
- Global_Sales - Total worldwide sales.