DATA ANALYSIS PORTFOLIO

Professional Background

I am an Analytical Chemistry, Advanced Diploma graduate from Durban, South Africa. I hold a certificate in Python programming from the Public Policy in Africa Initiative, Code Academy and also completed an Intro to Web Development course with Girl Code. Currently doing Data Analyst 1 course with Entry Level and also enrolled with the MTN App Academy, studying App Development.

I aspire to be a brilliant Data Analyst, as I enjoy the challenge of taking raw unorganized data and making it easy to interpret.

I believe in proactively chasing ones goals and being deliberate in the pursuit of seeing your dreams unfold through learning, practicing, applying yourself, asking for help and feedback and by simply just being curious and teachable.

To unwind I do gardening, crochet and hike.

Portfolio Outline

Udemy Project Description

Udemy, Inc. is an open online learning platform that has over 100 000 courses and a multitude of students. This project aims to give insightful solutions to the organization’s problem, brought about by the Head of Curriculum. The finally report to the CEO, would have satisfied the following objective:

- Ways to increase revenue in the next quarter.

The dataset from the manager contains four separate csv spreadsheets, namely: business studies, graphic design, musical instruments and web development. The spreadsheets comprised of identically layout data, that included these headings; course id, course title, url, price, number of subscribers, number of reviews, number of lectures, level, rating, content duration, published timestamp and subject.

The Problem

From the presented dataset, the problem of shortfall in revenue was a stumbling block in the compilation of the report. The questions below were asked to further gauge what the exactly is the business problem.

- What is the business problem?
- How long do you have to work on this project?
- What data should be collected to understand this problem? How should it be presented?
- What questions would you ask to better under the business problem?

To delve deeper into understanding the problem, we initiated a root cause analysis. The goals of this was to discover why a loss in profit is happening and what we can do about it. To uncover these answers a five step process was used which enabled us to understand the problem more completely. The following questions were asked and answered through that process:

- Q: “Why was there a decline in course sales?”
  A: The study only focused on four subjects.
- Q: “Why did the study focus on only four subjects?”
  A: To make it easier to track where shortfall is.
- Q: “Why make tracking the shortfall easier?”
  A: To be able to pin point which course(s) is responsible for the decline.
- Q: “Why pin point the responsible course(s)?”
  A: To see which ones to increase in price.
- Q: “Why increase the price?”
  A: To get more revenue.

We can deduce that the problem is really "the business" struggles to generate higher revenue. The revenue is generated from sales of course material. There are flaws in the way the business conducts its study on different subjects, which leads to wrong conclusions about the slump in course sales, hence the wrong pricing of their courses.

Data Design

The spreadsheets were consolidated, resulting in a single spreadsheet of 3677 rows. The data cleaning process included checking for duplicates, blanks, finding and replacing incorrectly entered data (in our case changing Subject: Web Development to Web Development) and headers were rewritten in a simplified and readable manner.

The Right/Left, IF and VLOOKUP Excel functions were used, to extract the date from the published timestamp column, to get how many courses are free and how many paid, to get a list of the top 20 most subscribed courses: their level, if there are free or paid, if there are any free beginner courses, duration and date published.

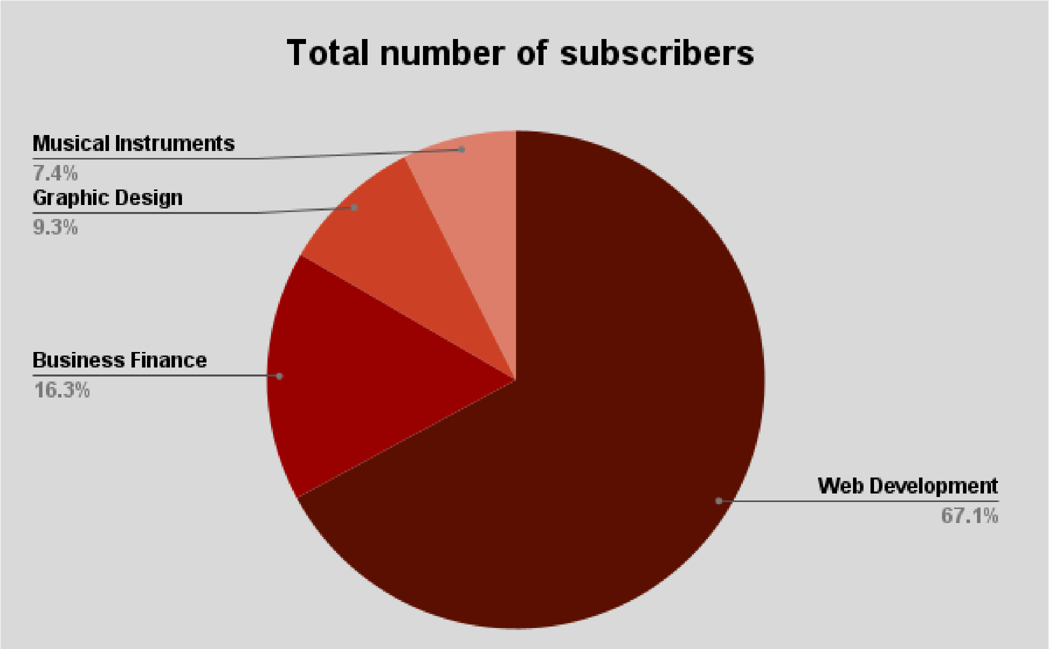
Pivot tables and charts of the total and average number of subscribers, average cost of subject, average content duration and average rating of subject were plotted.

For enhanced visualization, Tableau was employed to make two dashboards to graphically represent the data.

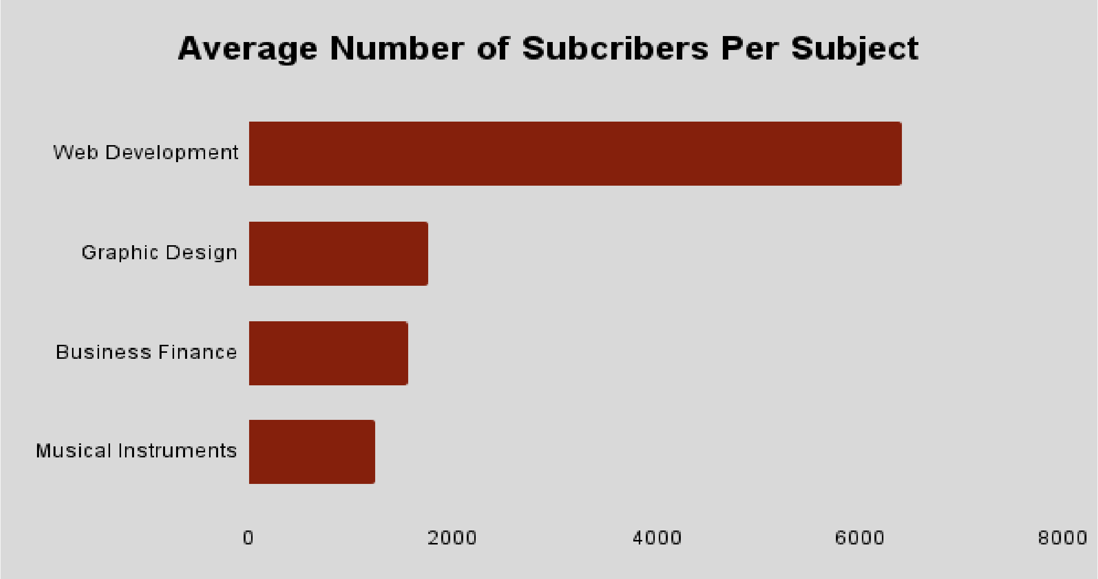
Findings

These are the insights were discovered from the visuals created:

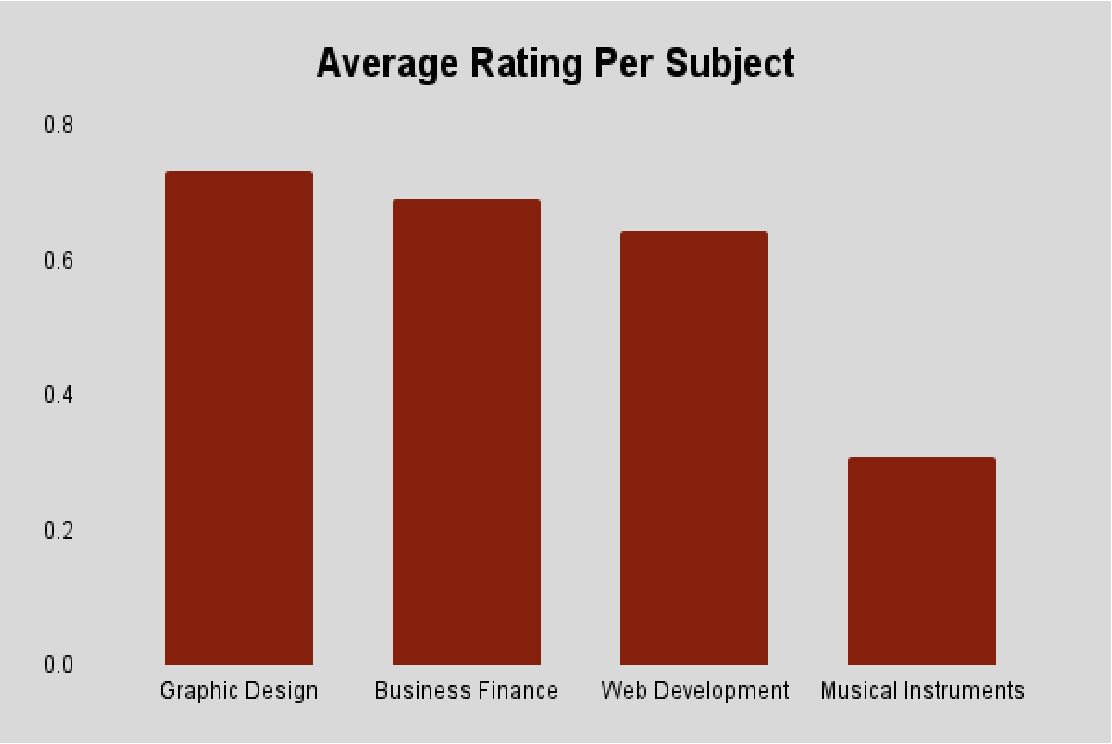
The pie chart below shows that Web Development had the most subscribers at **67.1 %** while Musical Instruments had the least numbers at **7.4%.**



Web Development had over **6000** average number of subscribers.



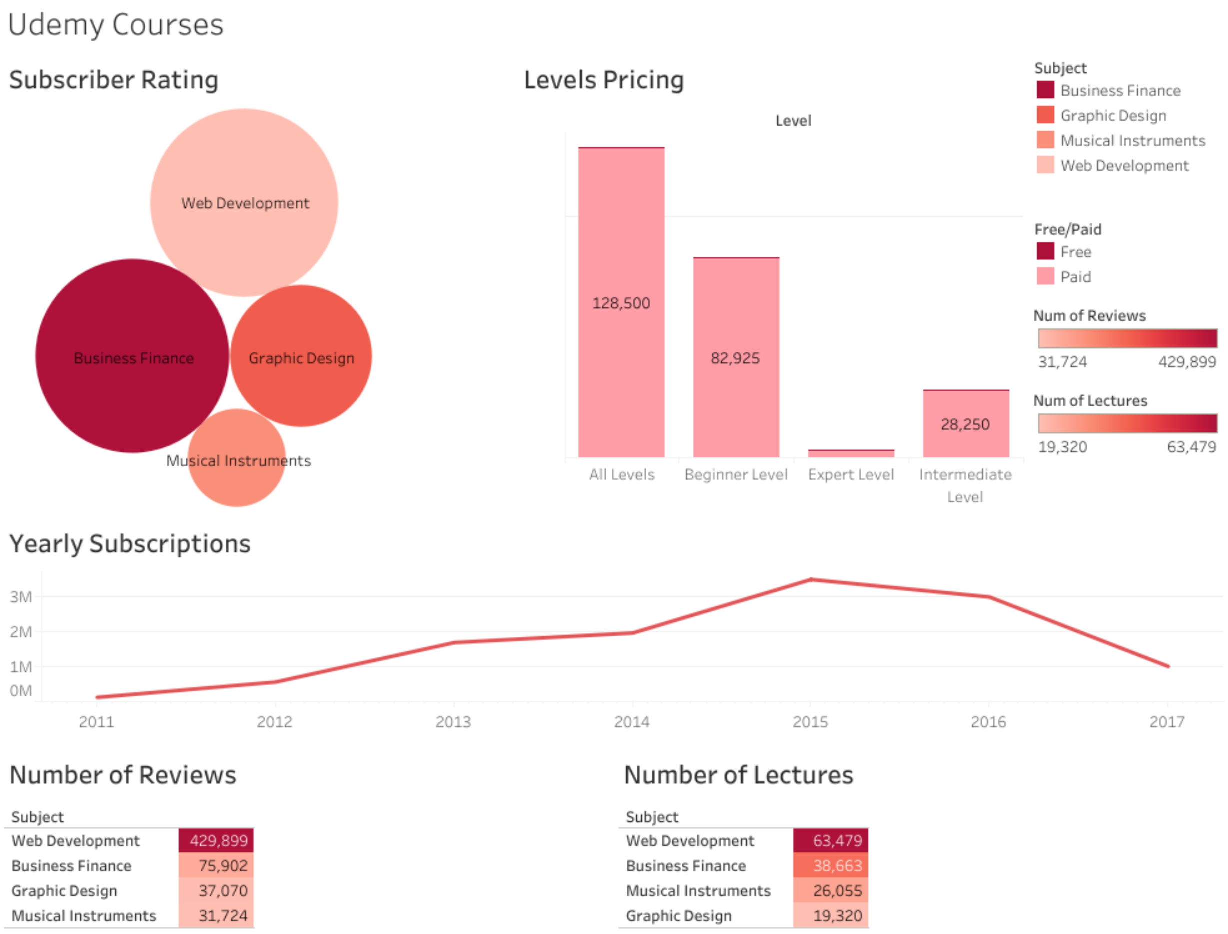
At ratings of **0.75** Graphic Design lead, with Business Finance closely following but Musical Instruments was less liked at **0.3.**



Finding 1

Using dashboards we see a clear representation of the tabulated data in visualization form. From the graphs, we deduced that in all the courses, overall Web Development as a subject far surpassed all the other three courses as it had more backing in terms of numbers.

In terms of pricing, beginner level courses ranked in more at **82 925**, while experts level only managed **3365** subscriptions.



Number of Reviews

Subject	
Web Development	429,899
Business Finance	75,902
Graphic Design	37,070
Musical Instruments	31,724

Number of Lectures

Subject	
Web Development	63,479
Business Finance	38,663
Musical Instruments	26,055
Graphic Design	19,320

Finding 2

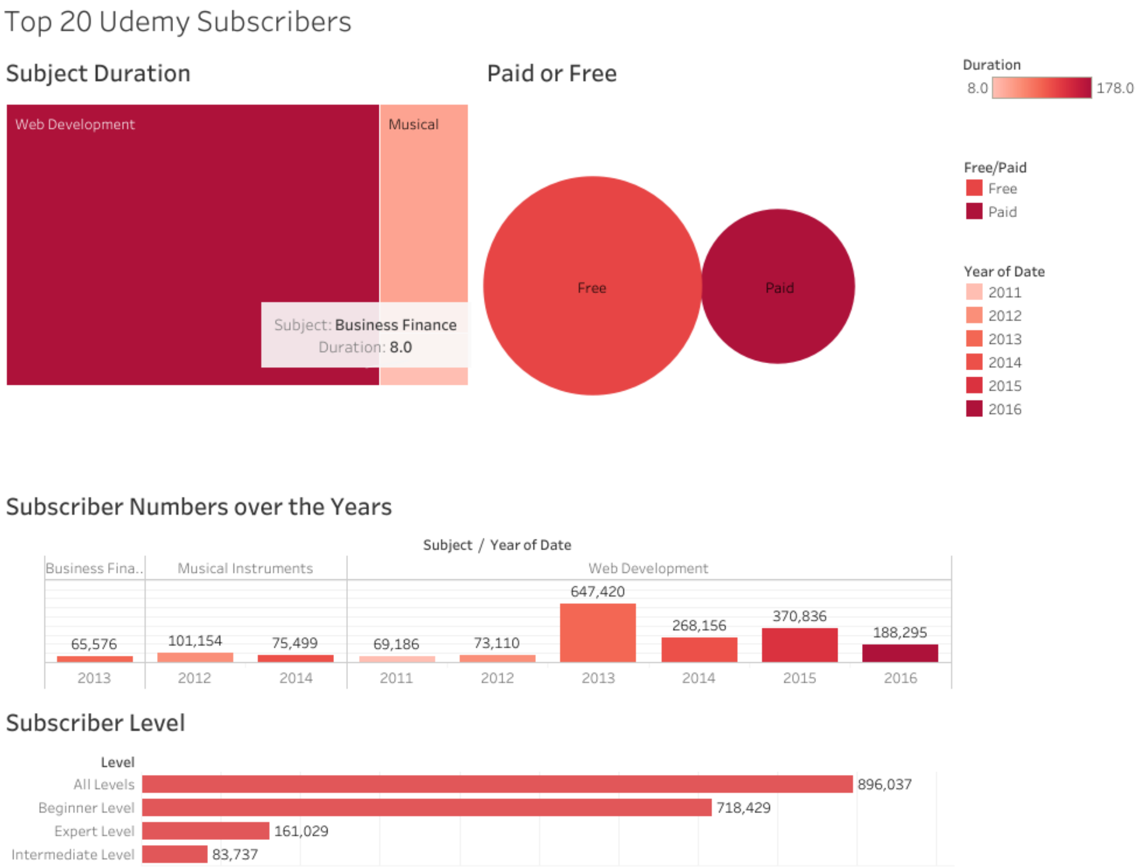
From the extracted top 20 courses, we found that at **268 923**, Learn HTML5 Programming from Scratch in the Web Development stream had the most subscribers.

Collectively, the content duration of Web Development was approximately **60%** more than the other subjects.

At **619 073** subscriptions, paid courses were about **50%** less than free courses.

The beginner level courses had more subscribers at **718 429.**

On average the duration of Web Development was at **178** and Business Finance at only **8.0**.



Analysis

To be able to track where shortfall is and which courses are causing the loss of revenue. The goal is to increase the price of the courses in order to match demand and profit.

There was a sharp increase in yearly subscriptions at **4M** in 2015, followed by a steady decline in the following years. The cause of the decrease could be attributed to the fact that free courses greatly outnumber paid courses, contributing to less income in those years.

However over the years, overall subscriber numbers, ratings and reviews have been in favor of the Web Development course. Boasting subscribers over **600 000** in 2013.

Conclusion

In summary, if you wish to take courses from Udemy, then you may want to look at web development which is the most popular course. It is worth noting that, more research need to be done to find out why the other courses had mediocre performance, perhaps look into the courses problem and attempt to amend the curriculum.

In order to increase revenue on certain courses, consideration may be done by reviewing the prices and perhaps change some or most of the free courses to paid as that may produce more revenue in return.

Furthermore to boast sales, more expert level courses could be introduced to increase overall numbers.

Capstone Project Description

The perovskite structure is a special class of oxides in the crystalline solid, their study is an area of interest in the scientific field because of their compositional flexibility, distortion due to the cation configuration and valence state electronic mixed structure. The data set was sourced from kaggle.com, crystal structure by Sayan Saha. It seeks to identify the perovskite structures of the compounds based on their spectral features. Each observation is described by 17 columns and 1 class column which identifies the oxide structure. Identification of 73 elements in the A and B sites of ABO3 structures, lead to perovskite oxides. To classify the crystal structure, characteristics of the elements, such as electronegativy, ionic radius, and valency and bond length were taken into account. Ideally, a perovskite has a crystal structure of ABX_3, A and B representing the two cations, with A (metal ion) being larger than B (metal or organic) in size and X (oxide or halide) is used as an anion while oxygen takes part in forming a cuboctahedron-like coordination environment around them.

The Problem

In this report we wanted to find out, how to determine the structure of crystals. So to answer this question we conducted a series of questions to give direction on what is really the problem presented from the dataset. Below are the questions;

- What is the problem?
- Did the manner of handling this study have an effect on the outcomes?
- How to predict the stability of a structure?
- What characteristics and properties play a part in crystal structure determination?

We did a root cause analysis, to best understand the problem. To find out how to classify perovkite crystals. The five whys were asked and answered as follows:

1. “Why is there an interest in perovkite materials?”
   They are efficient lost-cost energy materials.
2. Why are there lost- cost materials?”
   They have unique qualities.
3. “Why do they unique qualities?”
   They are structured differently from other crystals.
4. “Why are they structured differently”
   A: To aid in classifying them.
5. “Why classify them?”
   To be able to get which ones are best for what application.

We can assume that the problem is we want to know how to interpret the dataset, to be able to identify the desired perovkite structures. And to see what could be the main error or misunderstanding that can cause faulty findings.

Data Design

A single csv file with a dataset consists of 5330 ABO_3 perovskite-type oxides was collected. The data was cleaned using the Excel functionalities. This was done through sorting and clearing out empty cells. Missing data was accounted for, we removed duplicates and removed dashes.

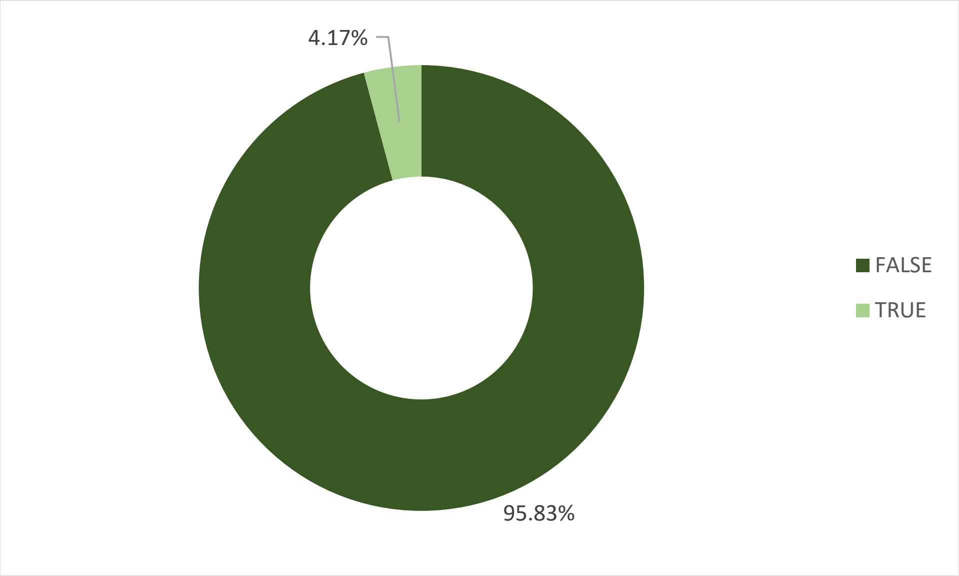
Took advantage of Excel’s tools, and did IF, VLOOKUP and left/right functions to organize and simplify the spreadsheet. Then we drew pivot tables were made to create charts.

To narrow down our data set, we used the 20 compounds that had the highest Goldschmidt’s tolerance factor, t and visualized the findings using Tableau dashboard for better presentation.

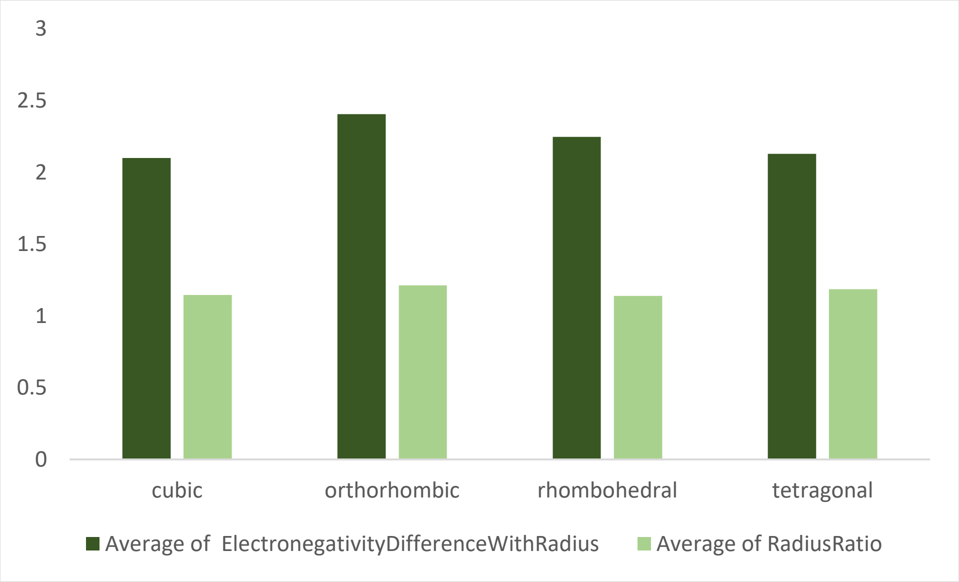
Findings

A common feature of the crystal structures of the perovskite oxides is that they have the same structure as Calcium Titanate ($CaTiO_3$). One characteristic that is responsible for the structure stability is the presence or absence of a large atom A at the center. Perovskite structures with orthorhombic, tetragonal or rhombohedral symmetry arise from the absence of such an element. In perovskite structure, distortions frequently occur due to deviation from ideal values of ionic size ratios between different A, B, O sites of the crystal. According to Pauling's rule for crystal structures, the critical radius ratio determines the allowed size of cation. At ratios of below 0.155, cation is too small, the compound will be unstable. Stability is established at radius ratios above 0.155. Cubic structure are the most common of all crystalline forms.

It is worth noting that only few crystal structures possess ferroelectric behavior with about **4.17 %** recorded in literature.



Electronegativity is inversely proportional to the size of atomic radius, evident in the column chart below, showing electronegativity difference values at **2.099 – 2.407** being higher than radius ratio values at **1.139 – 1.212**.

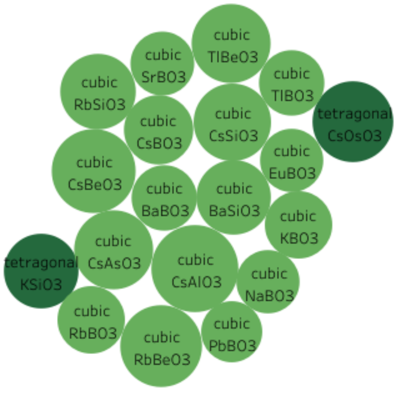


The evaluation of crystalline structures can be carried out using Goldschmidt's tolerance factor (T), to assess the geometric stability and distortions taking into consideration the ratios of ionic radii of A, B and X. Most of the structures were cubic because T values were in the range **1.177 – 1.389.** Lower tolerance factor values indicate higher symmetry, whereas higher tolerance factors indicate lower symmetry.

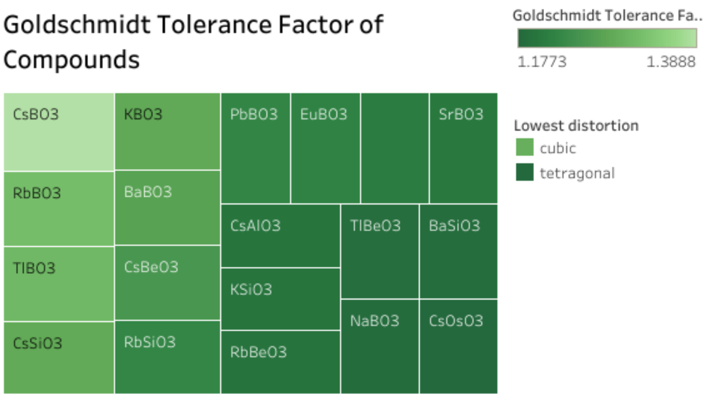
The comparison of average radius, bond length and ionic radius all showed that the crystals were in favor of cubic alignment.

Looking into the electronegativity differences of the crystals, we had acceptable values of **1.184 to 2.403**, all adhering to the ideal structure of perovskite with the except of $CsOsO_3$ and $KsiO_3$ which were tetragonal is symmetry.

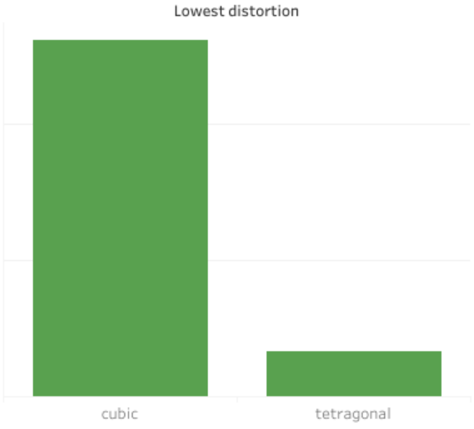
Structure of Compounds by Electronegativity Difference



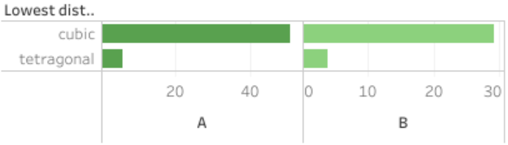
Goldschmidt Tolerance Factor of Compounds



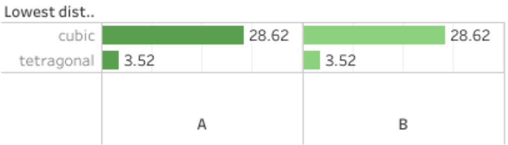
Average Radius by Structure



Bond Length Comparison



Ionic Radius Comparison



Analysis

The perovskites are important for their electron-transport properties, such as conductivity, paramagnetism and ferroelectricity.

For analysis purposes we focused on three characteristics of perovskites, that is, the mean bond length, electronegativity and Goldschmidt tolerance factor.

Mean bond length is used to describe the average distance between atoms in a covalent bond. It is dependent on several factors such as cation size and coordination number, distortions from ideal geometry to tetrahedral disorder. Bond lengths values range between 1-2 Å. The smaller the bond length, the greater the stability of the chemical compounds. Bond length is related to bond order, higher bond order indicate stronger bonds with greater changes in structure. Variations in mean bond length are examined in oxides, where many elements in the same period show similar trends due to their similar ground state electron configuration (i.e. the number of valence electrons). In oxides, mean bond length distorts because anions accommodate cation distortions by undergoing corresponding changes in coordination number and atomic radius. Each cation distortion results in a unique local environment around it, thereby influencing the next-nearest neighbor coordination and resulting in unique bond patterns.

Generally speaking, electronegativity is inversely proportional to bond length. This can be correlated with the fact that high electronegativity characterizes more stable bonds overall. Our compounds had moderate electronegativity, in accordance to Pauling's rule because the perovkites have are comprised of part metals, and metals tend to be less electronegative therefore affecting the whole structure.

Goldschmidt tolerance factor (T) is used to study and predict stability of different structures. For perovskites structures, T must be between 0.8 – 1.0. At values >1, A ions are too big or B is too small resulting in a tetragonal structure. At T, 0.9 – 1, A and B are ideal size, we have a cubic arrangement. At T, 0.71 – 0.9, the A ions are too small to fit into B ions interstices giving us orthorhombic/rhombohedral symmetry, however at $T < 0.71$, we have a different structure because A ions and B have similar ionic radii.

Conclusion

In summary, the characteristics of the perovskite materials been have used for the determination of the crystal structure. Cubic structures arise from the absence of a large atom A at the center. When the ratio of ionic radii moves away from the ideal value of T=1, there is a geometry strain and the crystals distorts. At lower tolerance factor values, the small sized atom A distorts to orthorhombic, tetragonal and rhombohedral symmetry however tetragonal symmetry prevails over orthorhombic symmetry.