Data Analysis Portfolio

The ability to scrape through data, clean the collected data in Excel and perform different analysis techniques, a keen eye for finding valuable insights and presenting them in an easy to understand way through Tableau, with extensive knowledge of SQL and Python while being an exceptional team player.

Professional Background

Recently graduated with a bachelor's degree in computer science engineering with a strong programming background in java, python and SQL. A highly driven individual with looking for a challenging position in an organization which will be the impetus to my professional and personal prosperity and will play to my strengths. Experience in product lifecycle management and field research as a part of my role in a local product based start-up. Have completed multiple data analytics projects using Microsoft Excel, SQL, Tableau and Python which have given me an all round experience in the major aspects of the role.

Table of contents

Udemy Project Description

Data on courses from various topics on the online education platform Udemy was provided to understand where opportunities to increase revenue may lie and to track the performance of these courses. The dataset included over six thousand courses with details like number of subscribers, duration, rating etc.

The dataset was reduced to twenty most performing courses on the platform.The objective is to perform analysis on these top courses and provide insights on how the revenue can be increased from these courses

.

Identifying the cause for the popularity of these courses was imperative as from that, we were able to determine the most plausible changes that can be made to the already existing system which will help in generating more income without hurting the already existing clientele.

After figuring out what made these courses so successful, we were able to think of various ways in which the sources of revenue can be increased greatly, like making 25% of the existing free of cost courses available for free and the rest of content being available behind the paywall of a nominal fee. On application of these proposed methods, we approximated almost double of the original revenue.

The Problem

Data on courses from different topics was given to us and we had to present data on course revenue in order to understand where opportunities to increase revenue may lie, and track the performance of courses. The manager suggested encouraging Web Development courses as they were the most famous on the platform.

The manager had to submit the report to the CEO in three weeks and that's why we had a time of two weeks to submit a complete report. Data that was required to generate adequate insights included course name, its cost, the number of people subscribed to it etc.

The problem was more complicated than just increasing the cost of existing courses as the popularity of those courses may drop down in some time as new hot topics emerge. If data on the growth percentage of a could have predicted which courses might become more popular.

Design

Data Cleaning – To make the data clutter free, we removed duplicates and empty data fields using in-built functions. To make the data more visually appealing, we changed the header of each row to be in the same format, small case separated by an underscore and changed the publish date from a specific time (ex. 2013-04-20T02:25:22Z) to a more standard format (ex. 2013-04-20) using the LEFT() function. One of the subjects was written in an incorrect format which was rectified by using find and replace function.
Refining the data – From this modified data, top twenty most subscribed courses were selected using the Large() function in MS-Excel or sortn() function in Google Sheets and this data was collected on a new sheet and the following corresponding data was derived using the VLOOKUP() function-

a) Their level

b) Whether they are paid or free

c) Whether they are a free beginner course

d) The date they were published

e) Their price

f) Their rating

Table 1: The final table definition

num_title	num_subscribers	level	course_fee	free_beginner_course	course_duration	date_published	price	rating

eg:

Quickstart AngularJS	64128	Beginner Level	Free	Free Beginner Course	1.5	2014-11-22	0	0.96

Findings

After we had the data that was required to figure out how revenue can be increased for the next quarter, we were able to look for patterns and analyze data presented in pie charts and bar graphs and we were able to figure out what makes these courses stand out from the rest of six thousand other courses available on the platform. On further analysis, we came to a conclusion that the best way of increasing the revenue was by increasing the number of subscribers and we were able to find various methods to double the number of people subscribed to these courses in order to drastically increase popularity and almost doubling the revenue generated for the next year.

Finding 1

To start off, we take a look at the number of people subscribed to the top twenty courses in a pie chart followed by a bar graph that depicts the percentage contributed by each course.

Figure 1: Subscribers for each course

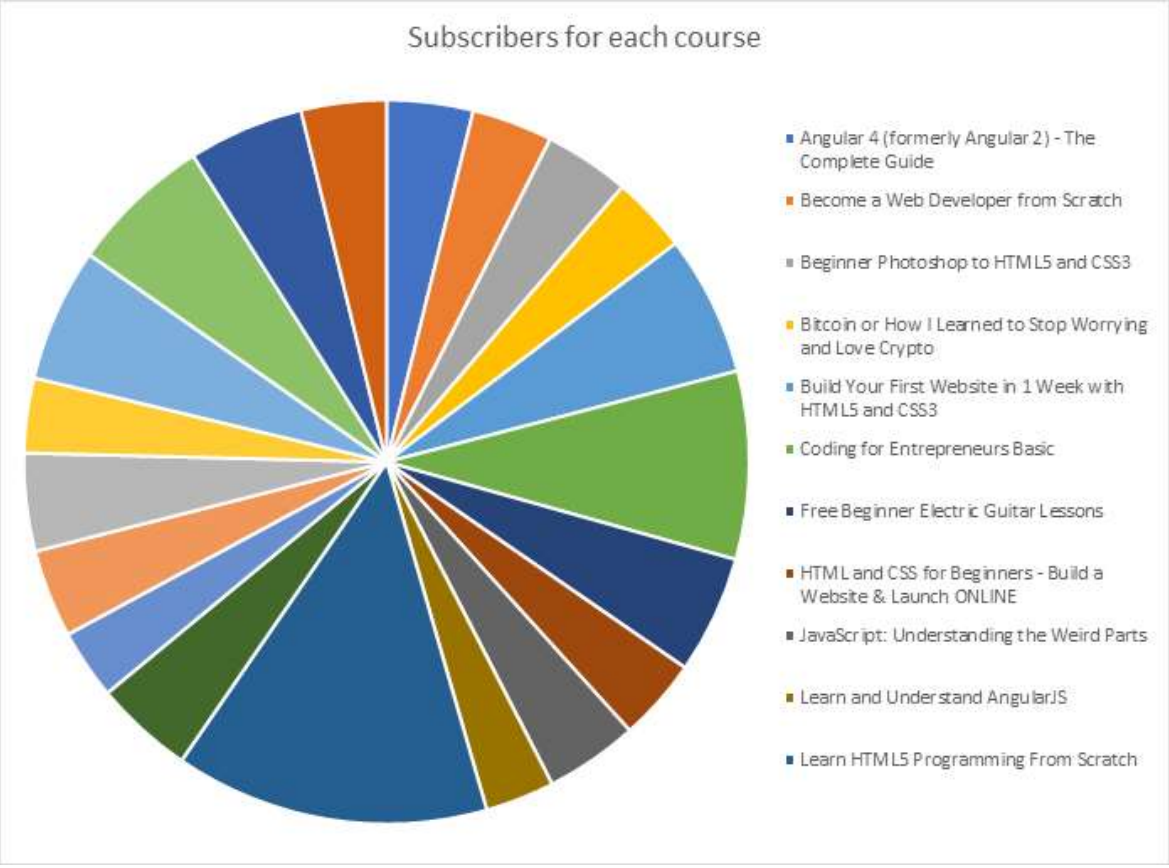
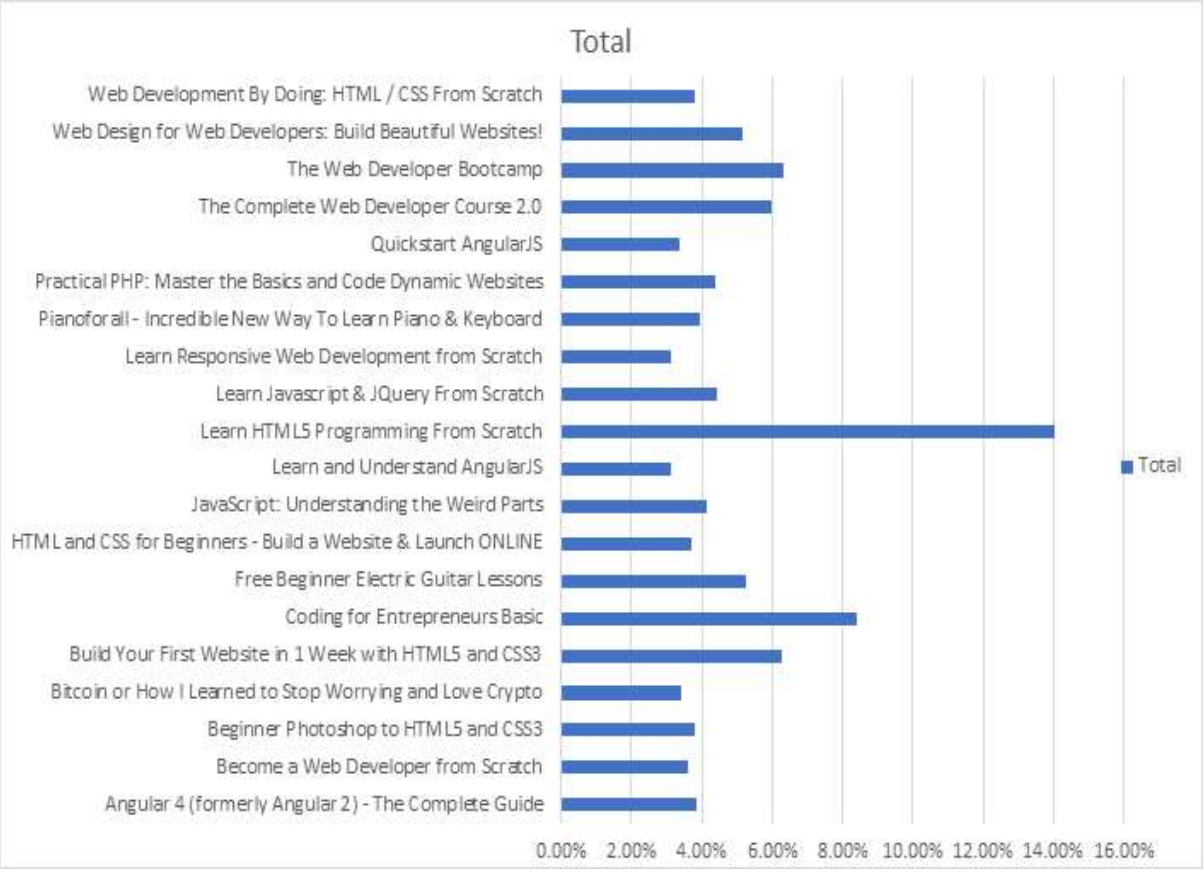


Figure 2: Percentage of total subscribers per course



From the pie chart and the bar graph it is clear that 'Learn HTML5 Programming….' Is the most famous course that is available on the platform with 14% of total subscribers for the top twenty courses. The most plausible reason as to why this is the case is because i) its beginner's level and ii) its free of cost.

Finding 2: Popularity of courses based on their length

The popularity of these courses also depend on their lengths. Most new users just want to experiment with online learning and they don't want to be tied down with a 40 hours course. Even some experienced folks will usually prefer courses whose duration is on the shorter side as can be seen by the following tableau visual -

Figure 3: Number of subscribers vs. Course duration

The attention span of an average human is 8 seconds which is just enough time for him/her to read the title and duration of the course and to decide whether he/she wants to go through it or not. As shown in figure 3, 'Learn HTML5….' has a whopping 268,923 subscribers and a duration of 10.5 hours whereas 'The web developer bootcamp' has less than half of that (121,584) as the length of this course is almost 4 times. Even though both these courses are related to web development, the course with shorter duration has a massive advantage.

Table 2: Average duration of the top 20 courses

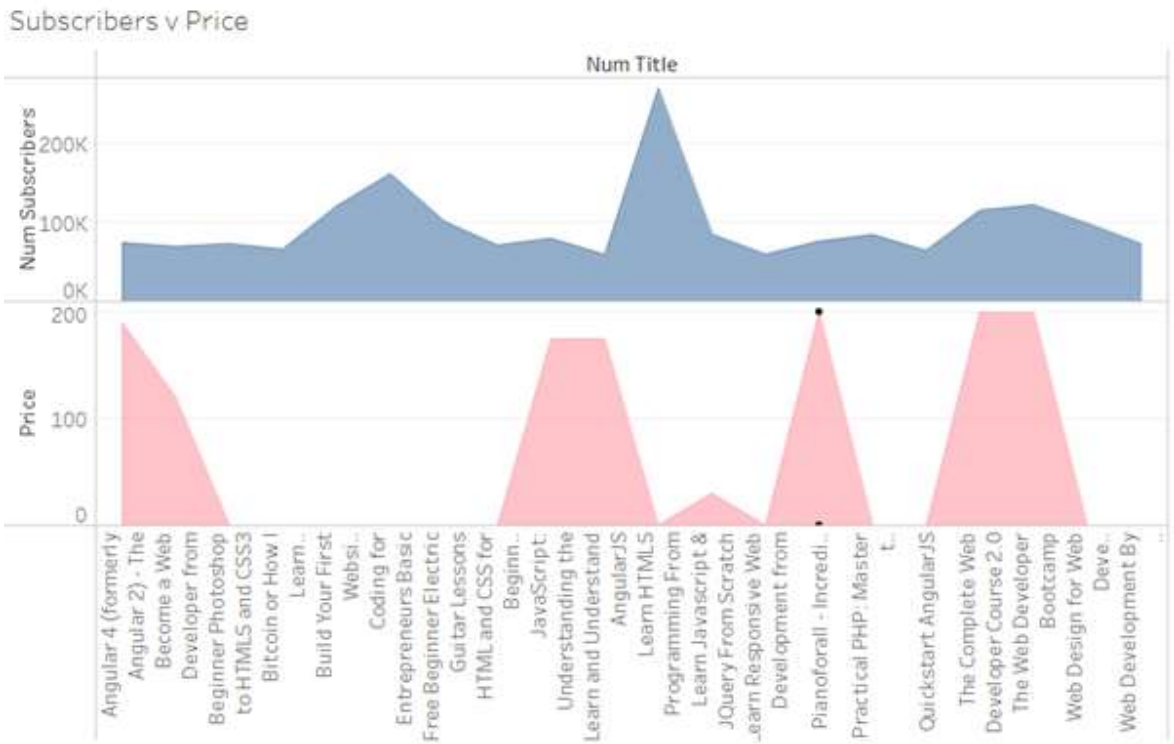
Name	Course duration
Angular 4 (formerly Angular 2) - The Complete Guide	22
Become a Web Developer from Scratch	27.5
Beginner Photoshop to HTML5 and CSS3	2
Bitcoin or How I Learned to Stop Worrying and Love Crypto	8
Build Your First Website in 1 Week with HTML5 and CSS3	3
Coding for Entrepreneurs Basic	3.5
Free Beginner Electric Guitar Lessons	4.5
HTML and CSS for Beginners - Build a Website & Launch ONLINE	6
JavaScript: Understanding the Weird Parts	11.5
Learn and Understand AngularJS	7
Learn HTML5 Programming from Scratch	10.5

Learn JavaScript & jQuery From Scratch	2
Learn Responsive Web Development from Scratch	4.5
Pianoforall - Incredible New Way to Learn Piano & Keyboard	30
Practical PHP: Master the Basics and Code Dynamic Websites	6.5
Quickstart AngularJS	1.5
The Complete Web Developer Course 2.0	30.5
The Web Developer Bootcamp	43
Web Design for Web Developers: Build Beautiful Websites!	3
Web Development by Doing: HTML / CSS from Scratch	1
Average Duration	11.375

The average course duration of these top twenty courses is 11.375 hours and therefore, existing courses which are way too long or too short and new courses should be targeted to be around 11-12 hours to appeal to more consumers.

Finding 3: The Revenue

Figure 4: Number of Subscribers vs Price



The above area graph is not really a surprise as it shows free courses are more popular than paid ones. Having said that, some courses stand out even though they are quite expensive. These courses might be so popular that the company could generate some extra revenue by slightly increasing the prices of these courses even further.

Table 3: Revenue generated by each course

Name	Price	Subscribers	Revenue Generated
Angular 4 (formerly Angular 2) - The Complete Guide	190	73783	14018770
Become a Web Developer from Scratch	120	69186	8302320
Beginner Photoshop to HTML5 and CSS3	0	73110	0
Bitcoin or How I Learned to Stop Worrying and Love Crypto	0	65576	0
Build Your First Website in 1 Week with HTML5 and CSS3	0	120291	0
Coding for Entrepreneurs Basic	0	161029	0
Free Beginner Electric Guitar Lessons	0	101154	0
HTML and CSS for Beginners - Build a Website & Launch ONLINE	0	70773	0
JavaScript: Understanding the Weird Parts	175	79612	13932100
Learn and Understand AngularJS	175	59361	10388175
Learn HTML5 Programming From Scratch	0	268923	0
Learn Javascript & JQuery From Scratch	30	84897	2546910
Learn Responsive Web Development from Scratch	0	59639	0
Pianoforall - Incredible New Way To Learn Piano & Keyboard	200	75499	15099800
Practical PHP: Master the Basics and Code Dynamic Websites	0	83737	0

Quickstart AngularJS	0	64128	0
The Complete Web Developer Course 2.0	200	114512	22902400
The Web Developer Bootcamp	200	121584	24316800
Web Design for Web Developers: Build Beautiful Websites!	0	98867	0
Web Development By Doing: HTML / CSS From Scratch	0	72932	0
Average	64.5	95929.65	**6187462.425**

As we can clearly see from the above table, the average revenue generated by the top 20 courses is ₹6,187,462.425 whereas the total revenue generated is ₹111,507,275.

Analysis

The main objective of this analysis was to increase the revenue generated by the twenty most popular courses offered by Udemy. Some of the graphs prove the integrity of the data by showing expected patterns. Following the 5 whys technique for root cause analysis, the first question that arises is *where lies the opportunity to generate more revenue?* And the answer to that can be either reducing the number of free courses offered by the platform as a lot of them made into the top twenty list or increasing the cost of the most popular courses. The second question that comes forth is *which course's price should be reconsidered?* which is answered by the analysis provided in this report.

Although increasing the price of every course may seem like an 'easy' answer, it can have a massive negative effect in the form of the number of overall subscribers going down substantially. Increasing the pricing must be done with great care and gradually to not off-put the customer.

The best way to increase revenue would be to push up the subscriber threshold required to make it into the 'top 20' which currently resides at 59361. By increasing the quality of the content provided in various courses of all levels and even slightly reducing the cost of some of the courses, we might be able to generate more revenue as the number of subscribers increases. Increased number of subscribers also means more popularity which is always a good thing. Over the period of next year, the company should work on increasing the threshold of the top 20 most subscribed courses to around 100,000 to generate a possible revenue of 200 million, almost double of what it is right now.

.

Conclusion

The business problem was pretty simple, to increase the existing revenue by analyzing the most popular courses available on the platform. The dataset provided was adequate enough but we could have generated some more valuable insights if we had more specific data like growth of subscribers in the last month or in the last quarter or both.
The solution was not quite as simple as the problem. The most easy method to generate more revenue would be by increasing the cost of the popular courses and reducing the number of free courses available but in reality, this could backfire pretty hard and might even result in overall loss. Another solution could be by re-evaluating the content provided by these courses and polishing them to be the best in the market but this plan has a ring of hit or miss to it, not to mention this would cost a lot of time and money among other resources.
The solution that was provided in this report was a bit more complex. Increasing the threshold of number of subscribers required to break into the top 20 is not only beneficial for marketing and spreading the word, but also risk free. The worst that can happen is not seeing an exponential growth in a short period of time which is easily overshadowed by the fact that this solution can make millions upon millions and work effectively in the long run.

Covid 19 Project Description

Covid-19 took the world by a surprise, everyone knew about the danger of such a highly contagious virus, but no one thought that it would affect every one of us to this scale and still, nearly 20 months since it made its first appearance, continue to force us to change our ways of life. Today, in this report, I'll use the dataset provided from John Hopkins University to dabble into some figures and see how the various countries have dealt with covid and possibly figure out when we can truly say we are covid free.

Disclaimer: The results presented in this report are based on just one database and might vary from the actual figures.

The problem

Pandemics are not really all that frequent. Factually, the last two great pandemics were spaced out by a century of time each and although Covid-19 wasn't quite as deadly as its predecessors, it certainly proved to be the biggest hurdle of all of them, mostly because of its timing. The world population is at an all time high which only means more targets for this highly contagious viral disease. The damage it has economically is so massive that it's going to take many years for us to bounce back from it. The problem is not quite as complex as the implications it has caused. With the help of some figures and a little bit of forecast analysis, we're going to look at the effect of covid and possibly answer the million dollar question, '*when is it going to end?'*

Design

The database has tons of data, with over 100,000 rows and around 50 columns of interesting figures but for this project we focus on 2 main topics: deaths due to covid and vaccination and its pace around the world.

After downloading the data, it was divided into 2 tables namely Covid Deaths and Covid Vaccinations to simplify the processing of the vast data. These two tables were then imported into SQL Database to perform complex queries. The VLOOKUP() function in excel is powerful but limited when it comes to choosing specific columns whereas SQL shines in these situations, therefore all the queries were performed in SQL and then the generated tables were imported into Excel for cleaning up.

Once in excel, necessary procedures were taken to ensure the data is ready to be visualized. All the cells with NULL were replaced with 0 as well all the duplicates were removed. The date in the tables is in the long format so they were converted into short date types to make sure the visualizations would work as expected. With cleaning of the data out of the way, the tables were imported into Tableau.

Table 4: Definition for table CovidDeaths

iso_code, continent, location, date, population, total_cases, new_cases, new_cases_smoothed, total_deaths, new_deaths

Table 5: Definition for table CovidVaccinations

iso_code, continent, location, date, new_tests, total_tests, total_tests_per_thousand, new_tests_per_thousand, new_tests_smoothed, new_tests_smoothed_per_thousand, positive_rate tests_per_case, tests_units, total_vaccinations, people_vaccinated, people_fully_vaccinated

Findings

Firstly, we're going to look at the emergence of covid and all the casualties it has caused.

FInding 1: The Effect

Table 6: Worldwide Statistics

Worldwide Statistics

Total Cases	Total Deaths	Death Percentage
19,78,15,444	42,15,197	2.13

These are truly some shocking numbers, over 4 million deaths over the period of last two years is a great loss to humanity but comparing it to the 1920 pandemic of *Spanish Flu* which caused around 40 million deaths, we can confidently say that the advancement of medical sciences and modern healthcare services have been more than helpful. Covid-19 is more contagious than it is deadly which is confirmed by the low death percentage of 2.13%.

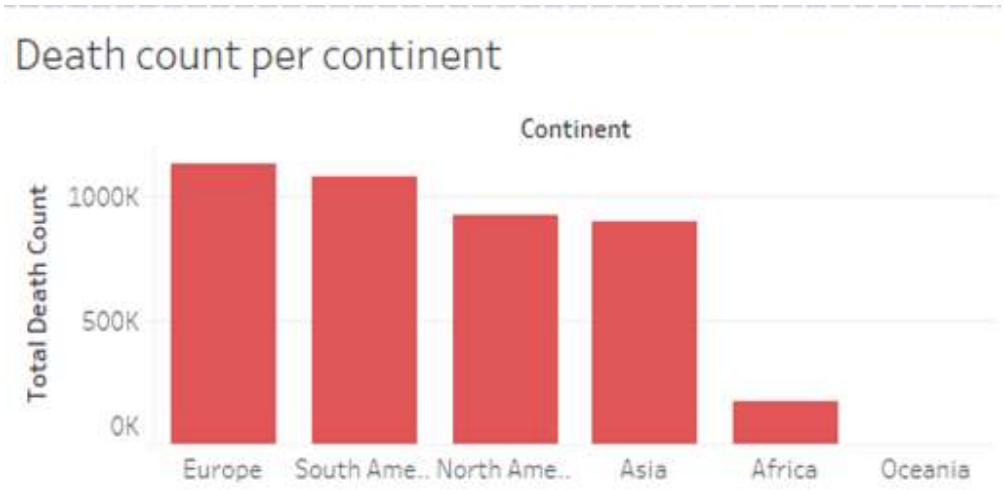
Now, let's break down the figures from the previous table

Table 7: Deaths per continent

Location	Total Death Count
Europe	1136947
South America	1082064
North America	924218
Asia	899588
Africa	170995
Oceania	1385

Some of these numbers are what someone would expect, whereas some numbers are quite surprising. Oceania is not only hardly populated, but also isolated from the rest of the world which makes the transmission of contagious diseases rather hard.

Figure 5: Deaths per continent



What's interesting to see is the positions of Asia and Europe in the above death count bar graph which is in descending order. The population density in Europe is 34/km^2 whereas in Asia it is 150/km^2. This development begs the question: what really happened in Europe that wrecked this havoc? One reason that comes to mind is because of how the European Union works. Easy travel between various countries inside Europe made these travelers a carrier of the deadly pathogen. Comparing this with Asia, the travel is not nearly as unobstructed. Another reason could be the differences in immunity of Europeans and Asians, with the latter being regularly exposed to humid conditions where diseases thrive, and the former being not used to this kind of virus. Maybe the massive outbreak in Europe could have been restrained if the travel restrictions were invoked sooner and maybe health organizations should research treating foreign pathogens as potentially hazardous to develop immunity in the people of such countries to prevent such a tragedy from happening in the future.

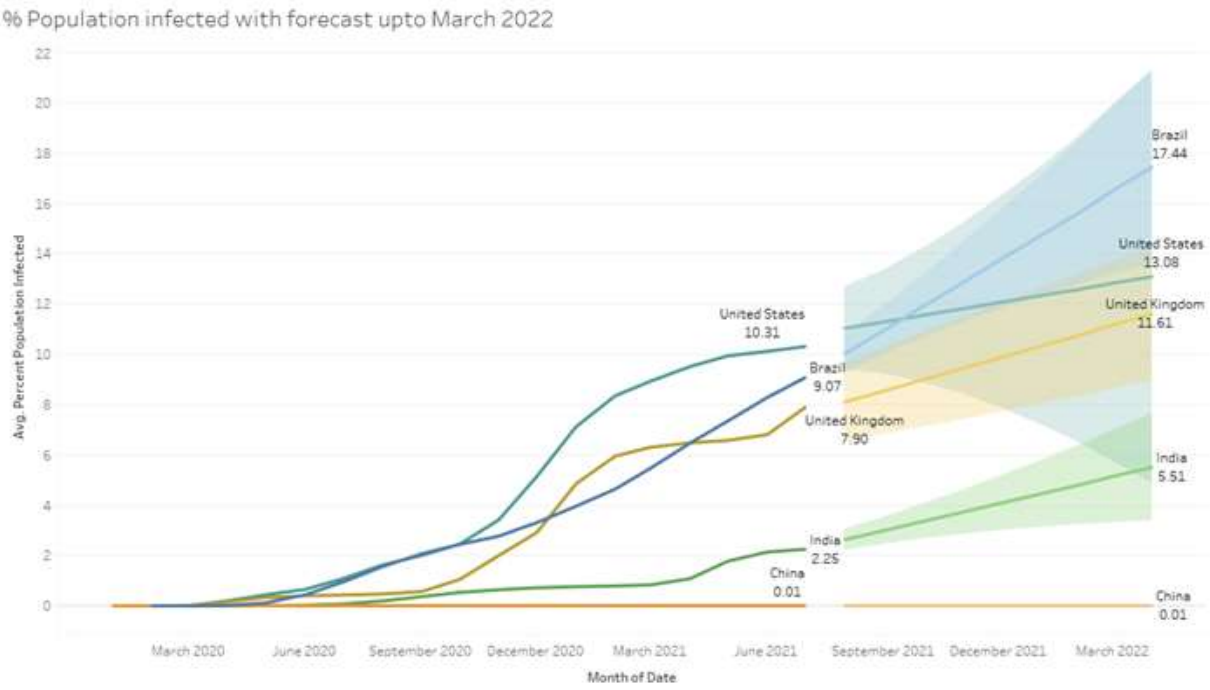
Figure 6: World Map of percent population infected per country

Continuing our drilldown, this world map shows how much different countries have been affected by covid individually. A deeper shade of red indicates more infection rate with small countries like Andorra and Seychelles having almost 19% infection rate. Granted these are small nations with minimal population, but the effect on their economy must have been devastating.

Figure 7: The effect of covid in India

Big countries like India will show lower infection rate but even 2.29 percent of its population gives results to a massive 32 million infections which paired with poverty and poor health infrastructure has caused great loss of life and economy.

Figure 8: Forecast Analysis using Tableau



The above graph indicates the infection rate of some of the world's worst covid-hit countries in July 2021 and the forecast up to May 2022. According to the forecast analysis, the future doesn't look very 'covid free' and countries like Brazil and the United States are going to have a rough start to 2022. The big changes in the curve for India in the month of April clearly shows there was a spike in infections, which marked the arrival of the deadly second wave.

Finding 2: The Only Viable Solution (for now)

Fortunately for us, vaccines were developed, and massive worldwide vaccination drives began in early 2021 with millions getting vaccinated everyday showing a glimmer of hope that there might be a day where we won't have to think twice before shaking someone's hand.

Table 8: Worldwide Vaccination Statistics

Worldwide Vaccination Statistics

Total Population	Fully Vaccinated	Percent Fully Vaccinated
7,764,825,504	600,095,416	7.73

Contrary to the popular belief, vaccines are not a cure, but they are quite effective in developing antigens in the patient's body. Most covid-19 vaccines are delivered in 2 doses over the period of 3-5 months and because of the time constraint while researching it, some people who are fully vaccinated were still being contracted by covid. But for the purpose of this report, we're going to assume that people who are fully vaccinated are safe.

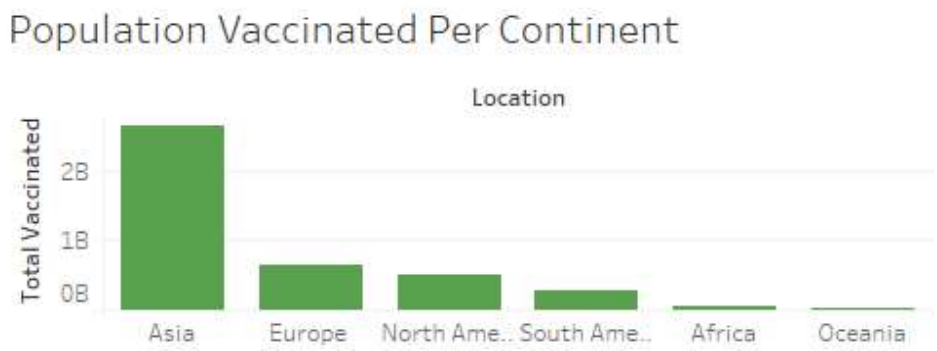
7.7% may not look like much but we must understand that vaccinating 600 million people was not an easy task given the short time frame. Thanks to vaccines, you can now go out to your favorite café or even catch a movie at the theatre if you're lucky. Businesses are opening again which will only boost the economy.

Table 9: Vaccine shots delivered per continent

Location	Total Vaccinated
Asia	2638298031
Europe	648160712
North America	497767219
South America	266561925
Africa	68249773
Oceania	15107353

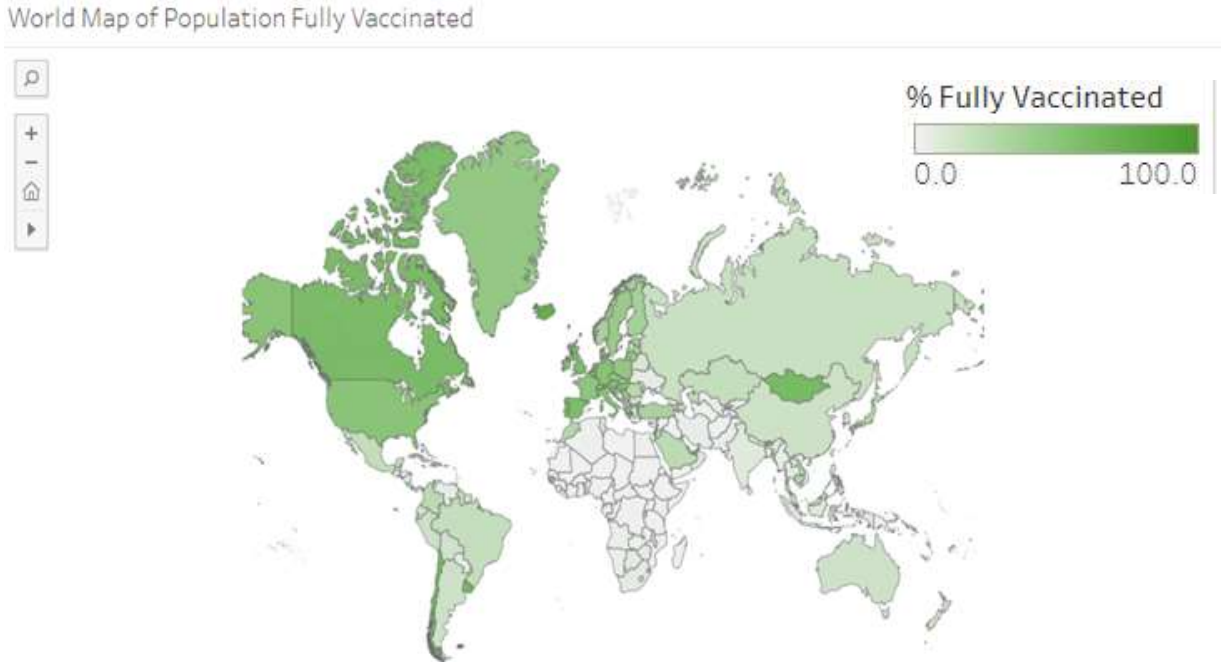
Bravo Asia! Over 2.6 billion vaccine shots is no common feat. Note that 2.6 billion is the total vaccine doses delivered, and not people fully vaccinated. Out of the 44 million people that live in Oceania, 15 million of the possible 88 million doses have been delivered.

Figure 9: Vaccination per Continent



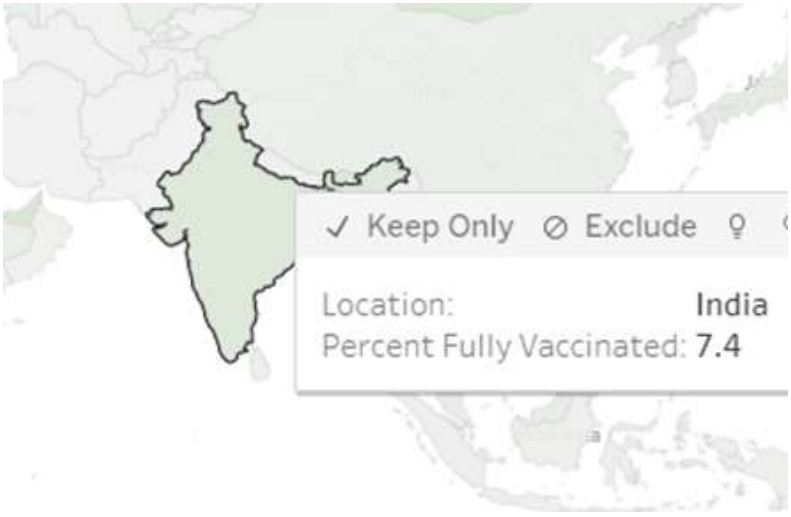
From this bar chart, it might look like continents other than Asia are slacking in their vaccination drives but in truth they are doing very well, maybe even better than Asia as they have much less population.

Figure 10: Percent population fully vaccinated per country



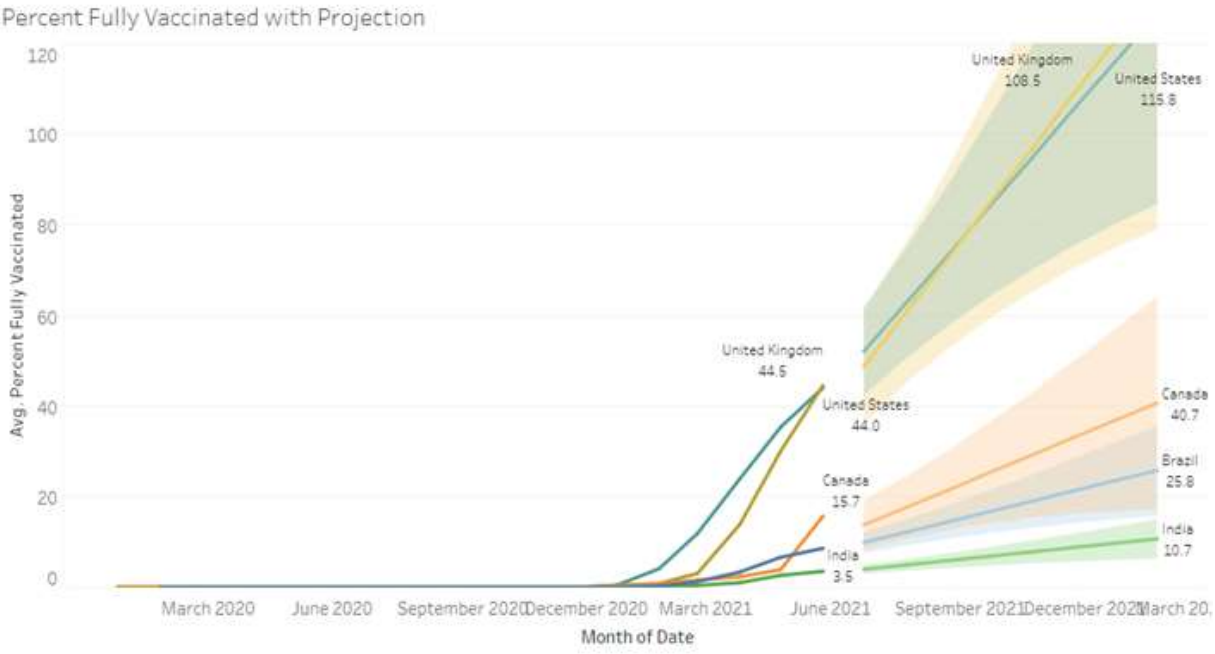
The above world map provides with great context about what's the current situation of each individual country with a deeper shade of green indicating that a country is close to vaccinating its entire population. Here we can see, even though Asia has a massive number of vaccinations, developed nations like the United States and United Kingdom are way ahead of countries like India in their quest for vaccinating their citizens.

Figure 11: Vaccination statistics in India



Vaccination drives are hard in a country with over 1.4 billion customers, but India faces the problem of illiteracy which acts as a roadblock in a smooth path of vaccination. Like India, many other Asian countries suffer from a high illiteracy rate which affects the % population fully vaccinated. Another obstruction is keeping up with high demand for vaccines. The production and shipment of the vaccines have to be done at such a large scale that it takes a toll on the efficiency of the whole procedure.

Figure 12: Forecast Analysis of vaccination drives in some countries



Conclusion

High community prevalence and slow confirmation of an outbreak after the first case was detected was a common cause of the high attack rate identified. Many of the positive cases were not identified quickly because they were asymptomatic or there was a lack of awareness in those interviewed of the wider spectrum of symptom presentation in older people. This resulted in testing not done in a timely manner. As a result additional control measures were put in place too late to stop the widespread transmission. Key to this is timely testing and reporting of results, in order that control measures can be put in place and so we must consider the system which may have created the optimal conditions for the virus to spread among the entirety of our population. The challenges with high community prevalence in the local areas, testing availability and turnaround times, combined with high occupant density, medical staff shortage indicators and the built environment risks re isolation or cohorting capability, placed everyone at risk of the swift spread of COVID-19.

Even after all the obstacles, hundreds of thousands are becoming safe from covid everyday through vaccination, and it shows in the form of the above graph. As of June 2020, 44% of the U.K. and U.S. population have been fully vaccinated and the projection shows that by November of this year, they will have completed their vaccination drive. It is important to note that the forecast analysis that's being used here is based on the data from the last 12 months and since that is scarce in the vaccination department, the analysis may be a bit off. Nonetheless, it is evident from the decisions of world leaders to reopen the world that they believe that the worst is behind us. Scientists from all over the world have been warning us about the emergence of a 'third wave' because of all the post-lockdown activity and the threat is real and I'm sure that by now everyone is aware of the potential damage covid can cause to not only an individual, but also to the world economy. Keeping this thought in mind, I think we can start to look ahead to a time where medical staff can finally take a breather and all the struggling businesses can sell their products or lend their services. It might not be in 2021 for everyone, but we've made it through the hardest times and only the final stretch remains before we can finally let go and go back to life before covid or whatever the new normal might be.