**PROBABILITY AND STATISTICS PROJECT REPORT**

You can find answers to each of the points/parts mentioned in the project under the **Point No:** headings.

**Point 1: Covid-19 Data of Pakistan**

|  |  |
| --- | --- |
|  |  |
| **Confirmed Cases** | **924,667** |
| **Deaths** | **20,930** |
| **Recovered Cases** | **848,685** |
| **Total tests** | **13,316,397** |
| **Active Cases** | **55,052** |
| **Critical Cases** | **3,767** |

**Link:**

<https://www.worldometers.info/coronavirus/?utm_campaign=homeAdUOA?Si#countries>

**Interpretation:**

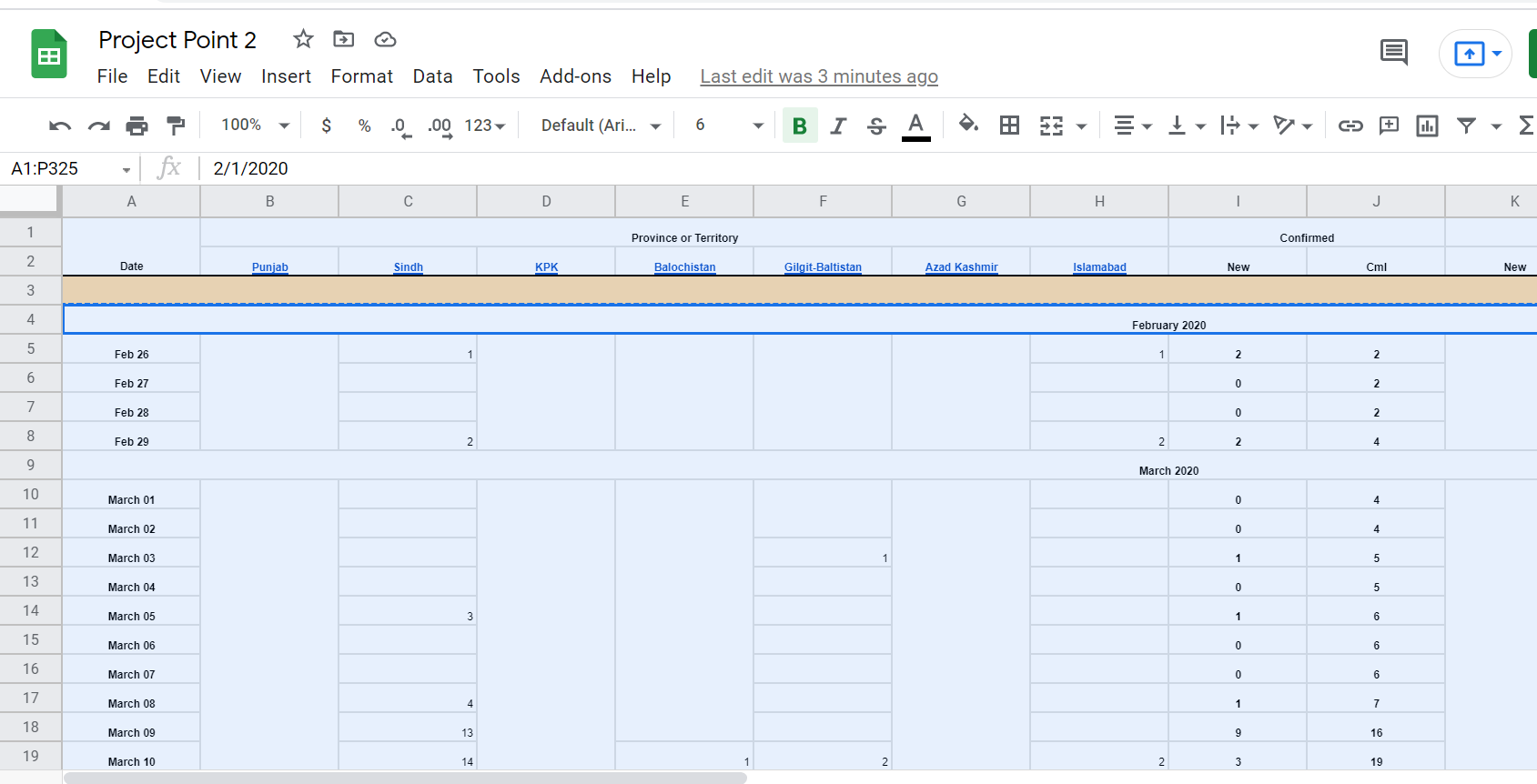
From this table, we can observe that the total number of Covid tests or PCR tests done is 13,316,397, out of which 924,667 came back positive, which is about 6.94 percent of the total tests. Then from the total positive cases, 869615 cases had an outcome, with deaths being 20930 and the number of recovered cases being 848685. Calculating the percentages of deaths and recovered cases, we get 2.26 percent for deaths and 91.78 percent for recovered cases. This leaves us with 55052 active cases, out of which 3767 cases are critical cases which accounts for 6.84 percent of active cases.

**Point 2:** Summary of the total/confirmed cases province wise

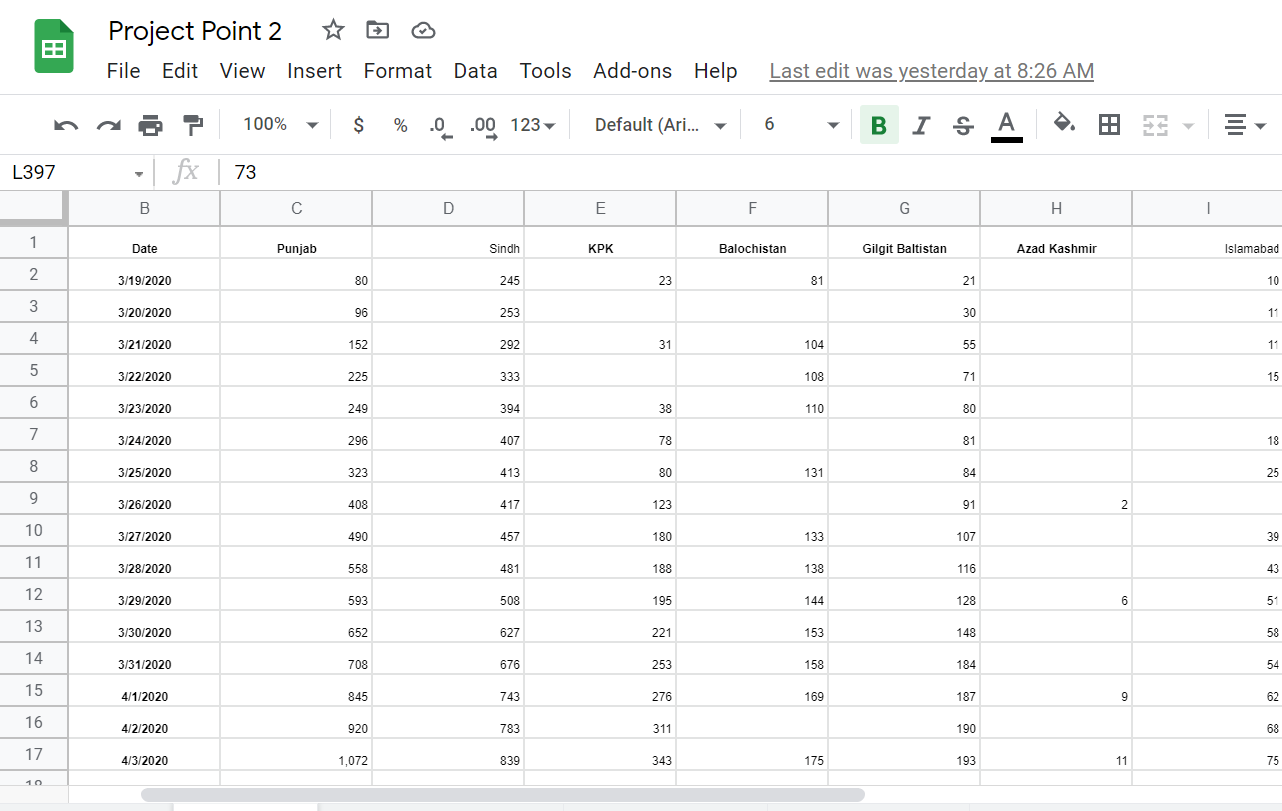
Link provided for data: <https://en.wikipedia.org/wiki/Template:COVID-19_pandemic_data/Pakistan_medical_cases>

Step 1) Imported Data from given link into spreadsheet.

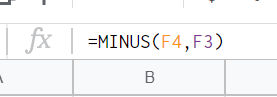
(Note: I was only able to Import Data from date 3/19/2020 to 4/29/2021 due to the large size of data and problems while copying from wikipedia to spreadsheet)



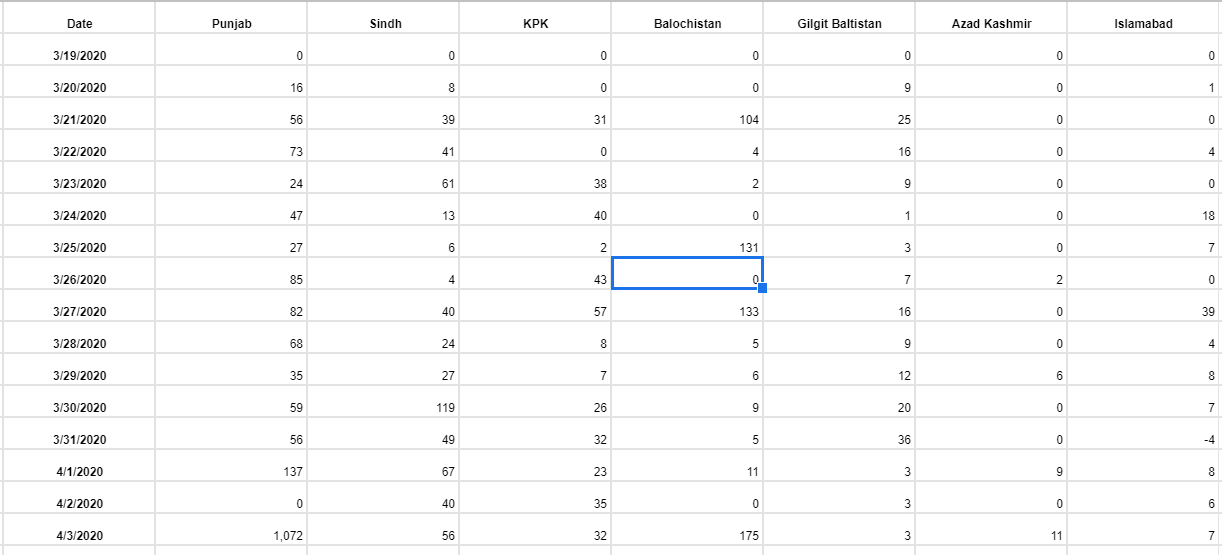
Step 2) Simplified the data by deleting excess columns and rows (NEXT PAGE)



Step 3) Data was provided cumulatively meaning the sum of total cases till date was given at each date. In order to get only the cases on each day, I subtracted the cases on the previous day from the current day for each day using:



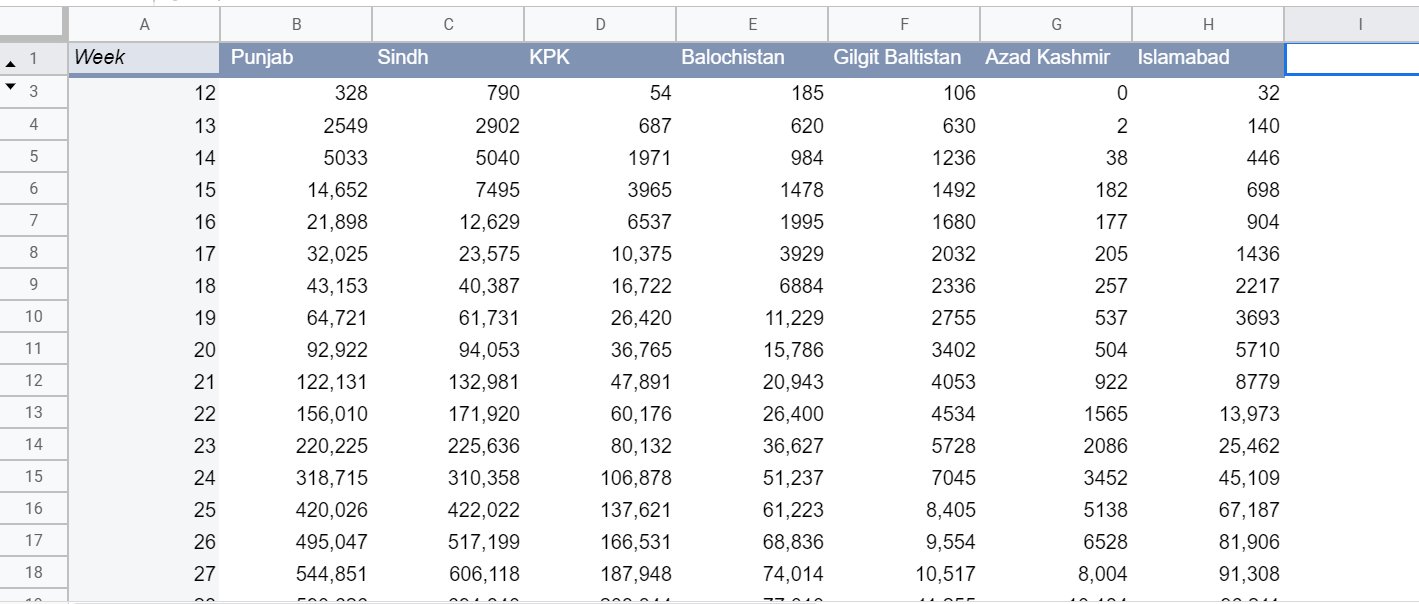
After that I had each individual date’s cases not total cases till each date. (Look at image below)



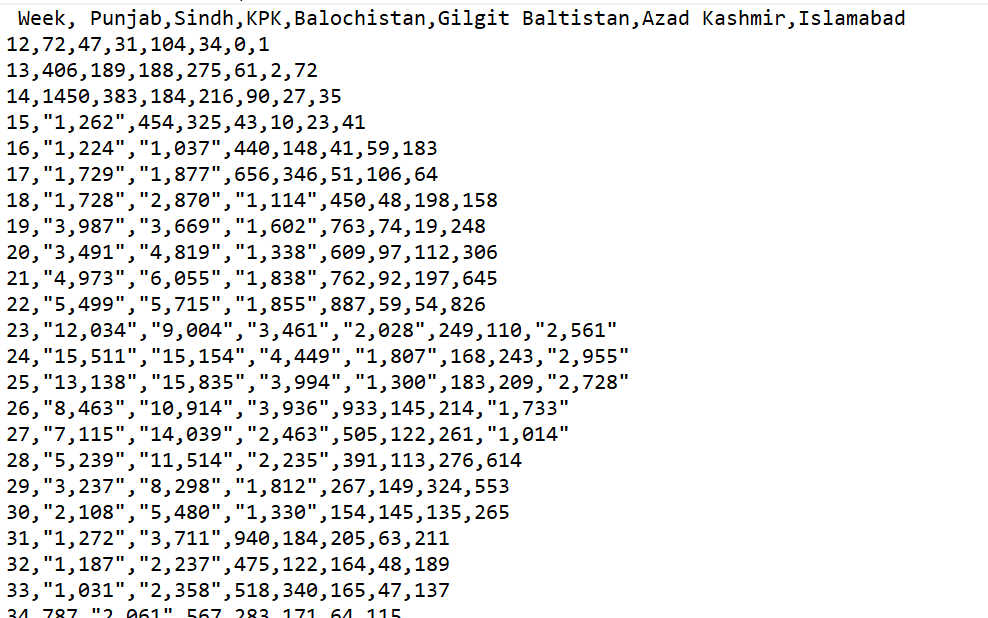
Step 4) Organized cases according to weeks instead of single days(Dates) using command

ArrayFormula(weeknum(B2:B)) (Basically gives the week numbers for each date)

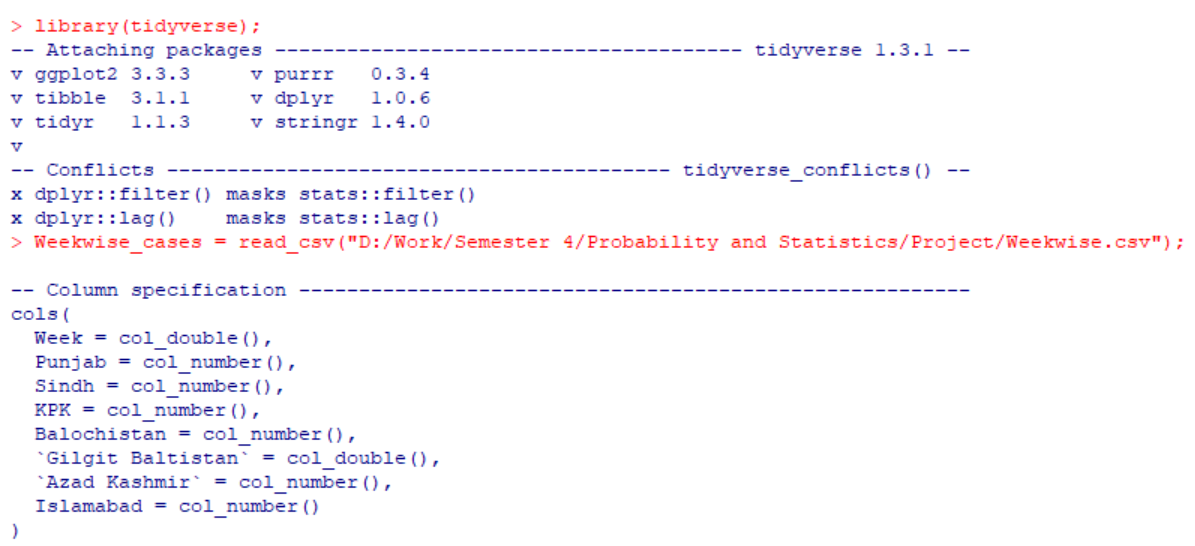
(Note: Week Number starts from 12 (3/19/2020))



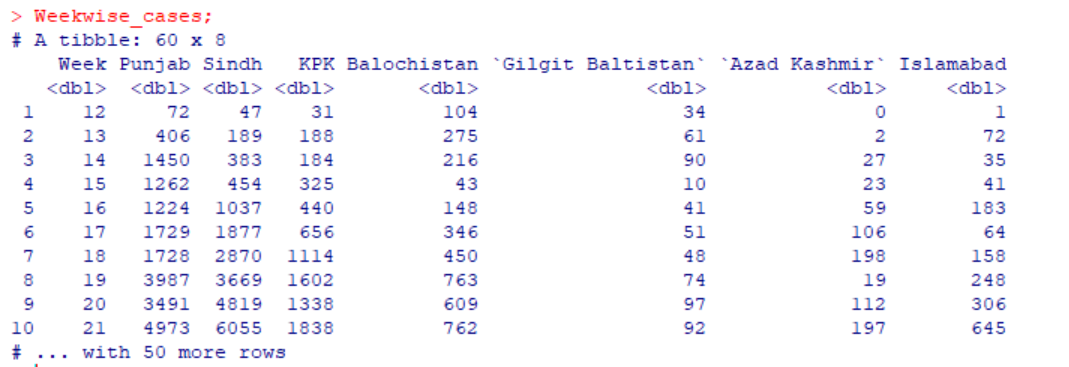
Step 5) Converted Data into .csv file



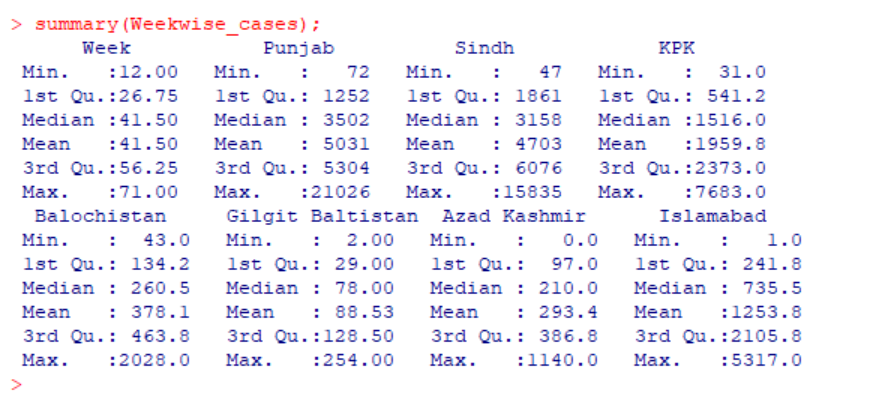
Step 6) Using tidyverse library in R and its read\_csv() function, read the data from the CSV file into an R dataframe



The data was now loaded into an R dataframe



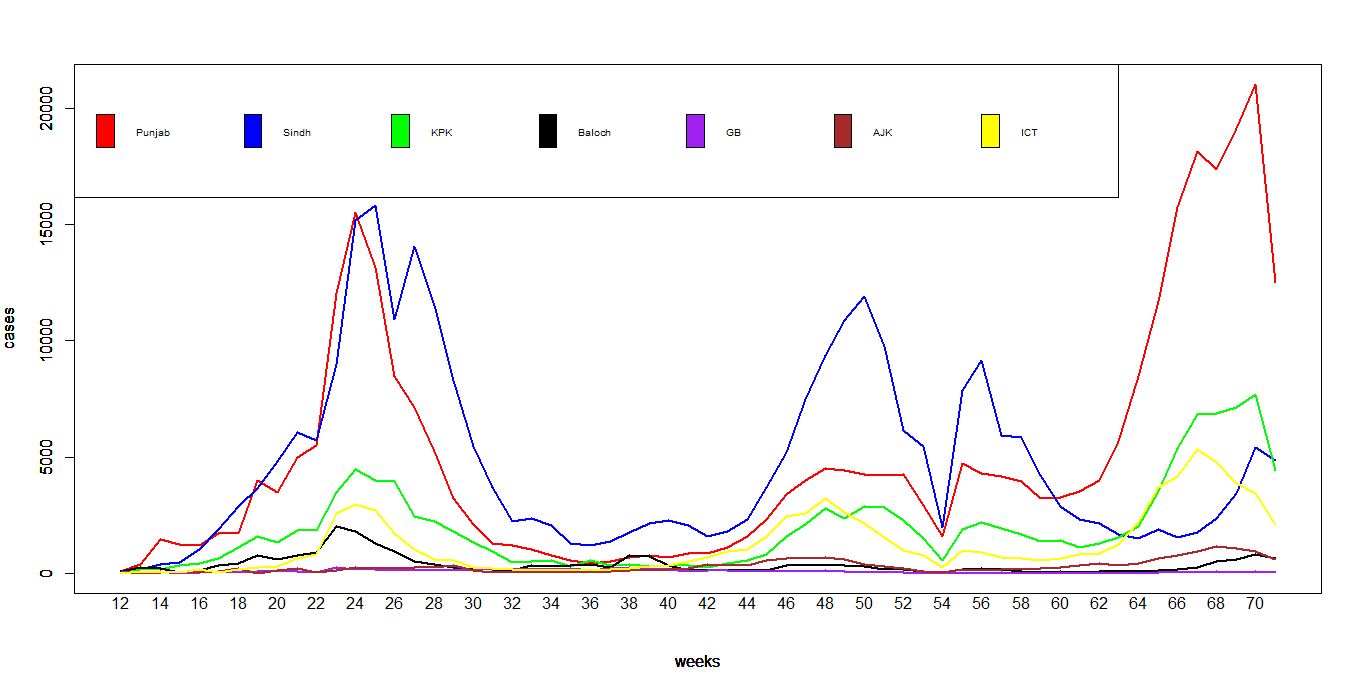
Step 7(final)) **Summary** of each province using summary() command in R.



**Interpretation:**

The summary basically shows the minimum, maximum, mean, median, 1st and 3rd quartiles of the cases of each province **per week.** The summary shows that Balochistan, Azad Kashmir and Gilgit Baltistan’s **mean** cases **per week** were on the low end (378.1, 88.53, 293.4 respectively) and Punjab and Sindh’s **mean** cases **per week** were on the high end (5031, 4073 respectively). Punjab had the **most** cases **in a week** out of all provinces (21,026), followed by Sindh(15835). Islamabad and Azad Kashmir had the **least** cases **in a week** (1 and 0 respectively). In addition, the 1st to 3rd quartiles **of cases per week** are also provided for each province.

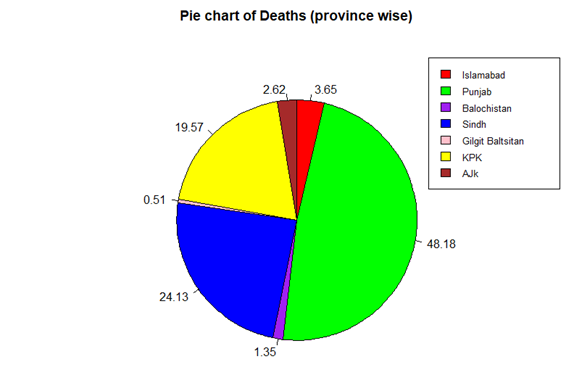
Here’s an R line graph that compares the week wise cases for all provinces using the data used to generate the summary:

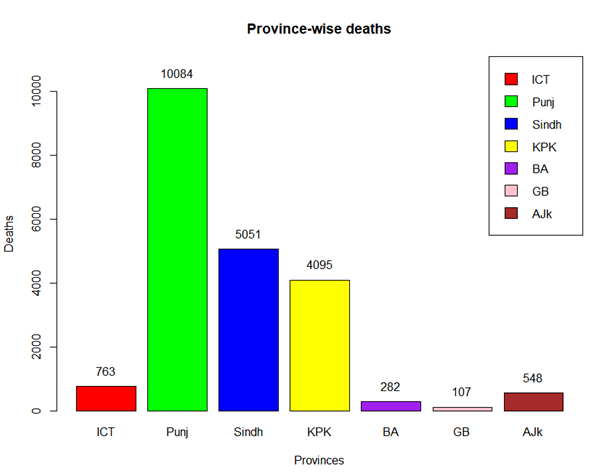
Link to spreadsheet used for this Part: (Please have a look) https://docs.google.com/spreadsheets/d/1Sh3L0fqVpagBbQ3gXKMGwC3vy\_yKTilPGqC9ackZNDE/edit?usp=sharing

**Point 3:** **Deaths** (By Province):

|  |  |
| --- | --- |
|  | **Deaths** |
| **Islamabad** | 763 |
| **Punjab** | 10,084 |
| **Sindh** | 5,051 |
| **KPK** | 4,095 |
| **Balochistan** | 282 |
| **Gilgit Baltistan** | 107 |
| **AJK** | 548 |

**Link:**

<https://covid.gov.pk/stats/pakistan>



**Interpretation:**

From the bar Chart we can observe that Punjab has the most deaths out of all the provinces with 10084. This may be due to the fact that Punjab is the most populous of all other provinces. Out of the total 20930 deaths, Punjab accounts for almost 48.18 percent of total deaths. Coming in at second position is Sindh province with 5051 deaths which amount for 24.13 percent of total deaths. Third, fourth and fifth belong to Islamabad with 763, Azaad Jammu Kashmir with 548 and Balochistan with282 deaths. The least amount of deaths were reported in Gilgit baltistan with 107 deaths, which is only 0.51 percent of the total deaths.

**Point 4: The survey**

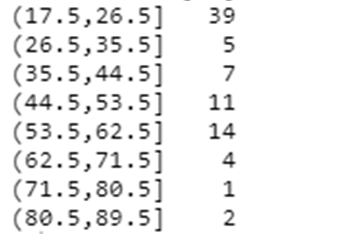
Link to survey form: <https://forms.gle/phvJHafG7oR4SWJT6>

Total Responses gotten: 88 total (83 by the time of doing Point 5)

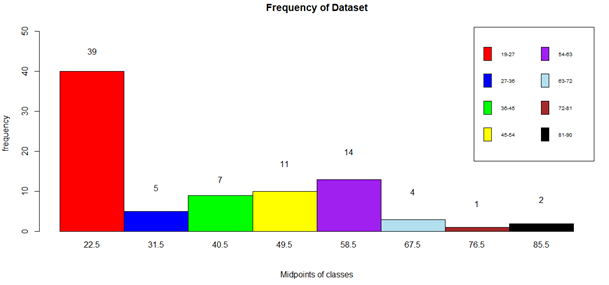
**Point 5: The Freq distribution and graphs of vaccinated population (using survey)**

Number of classes: 8

Class Width: 9

Classes Frequency

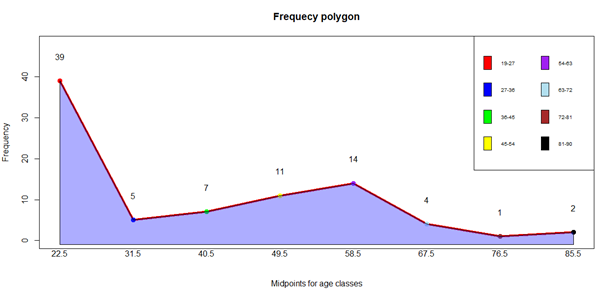
**Histogram:**



Interpretation for histogram:

At the time of making these graphs, we got a total of 83 submissions. We divided the data into 8 classes, each with the class width of 9 elements. From the histogram, we observe that the most amount of responses i.e. 39 responses came from the age class of 18 to 27. In contrast to this, the least amount of responses were submitted from the age class of 72-81 with just one response.

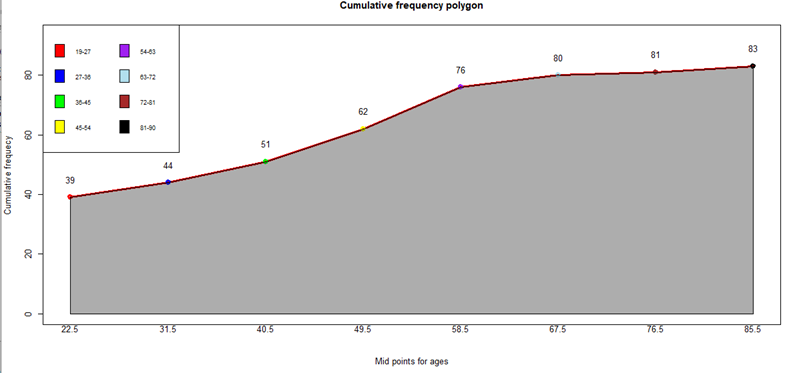
**Frequency Polygon:**



Interpretation:

By looking at the frequency polygon graph, we can see that there is no distinct pattern with most people in the 18-17 group. The number of people in other groups are far lower than the 18-27 age group. If we try to find the average number of people in each group, it turns out to be just over 10.

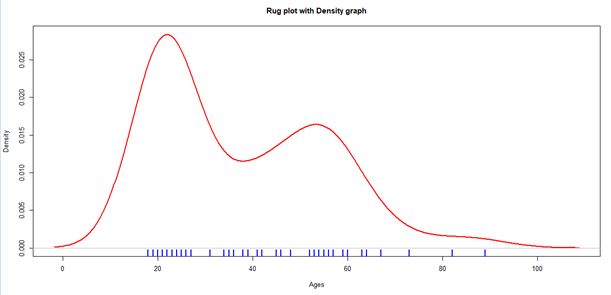
**Cumulative Frequency Polygon:**



Interpretation:

Cumulative frequency polygon provides us with the information about the total number of submissions which we can see displayed over the age class of 81 to 90, which is displayed by a black dot in the graph. From the legend displayed on the top left of the graph, we can get information about the different classes and the colors with which the age groups are displayed. .

**Rug Plot:**

****

Interpretation:

In this graph, rug graph and density curve graph is displayed. Density curve is displayed with a red line and the rug plot is displayed with small blue lines along the x axis. We can observe that age density is highest near the 18 to about 30 ages and then density graph hits second highest near the 50 to 60 ages.

**Point 6: Comparing Death Rate of Pakistan with other Countries**

Source of Data:

<https://www.worldometers.info/coronavirus/?utm_campaign=homeAdvegas1?%22%20%5Cl%20%22countries#countries>

I took the total cases and total deaths for several countries and inserted them into the data frame. For the death rate column, I used the formula:

Death Rate: Total deaths/ Total Cases

I inserted death rate column into the dataframe using:

*Data$Death\_Rate = Data$Deaths/Data$Cases;*

Total Cases, Deaths, Death Rates of several Countries including Pakistan.

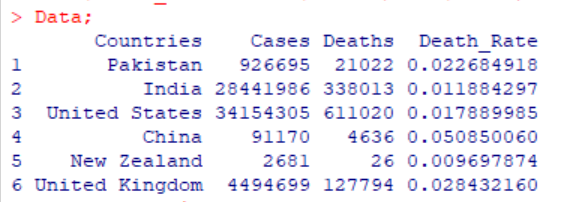


Fig 6-a

Visualized with a horizontal Barplot:

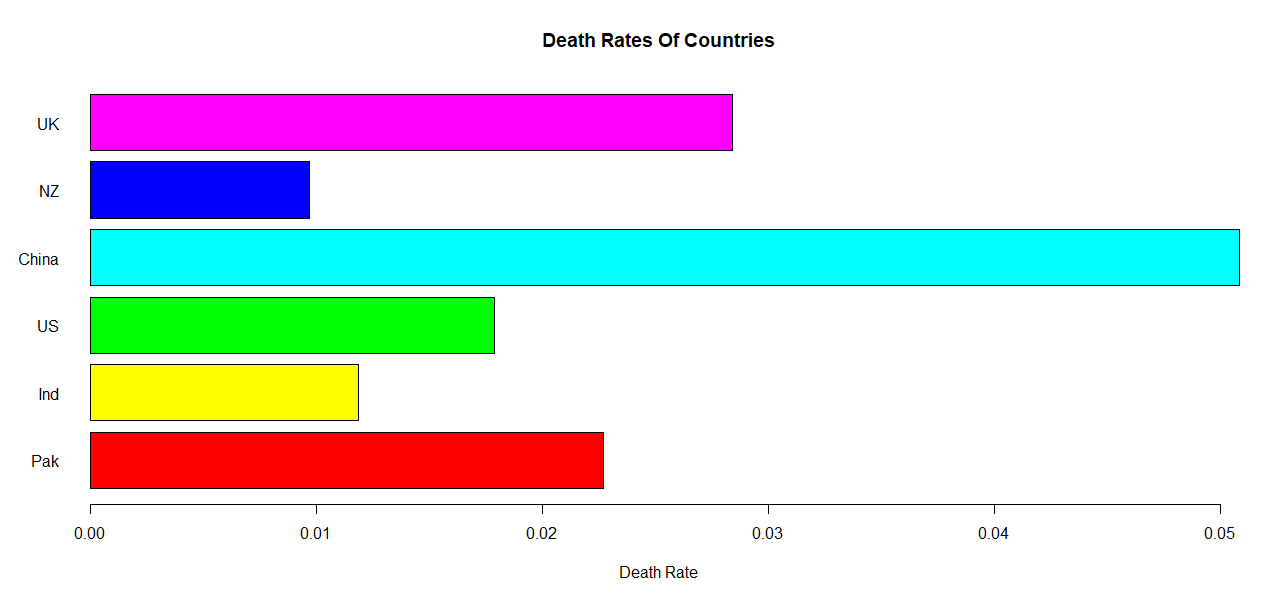


Fig 6-b

**Interpretation**: It can be observed from the barplot that China has **by far** the **highest** death rate. (Although it can be observed in Fig 6-a that the number of cases in China is amongst the lowest in the list of given countries). The death rate of all the countries given is less than or equal to 5% (China’s being roughly 5%). Pakistan’s Death Rate falls right in the **middle** of the pack at around 2.3%. US and IND despite having the highest number of cases by far, have relatively lower death rates than Pakistan (at round 1.7% and 1.1% respectively). NZ however beats all the countries, having the lowest death rate at 0.9% with the lowest number of cases as well (see fig 6-a).