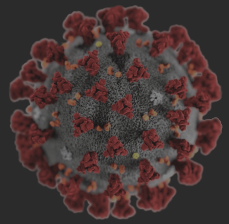


COVID-19 Analysis

Reston Big Data Batch

https://github.com/revaturelabs/201005-reston-bigdata.git

Technologies

Data Acquisition

Tanner Hall, Sean Horner, Timothy Mickle, Samuel Owens

Data Sources

- COVID-19 Case Reporting from the CDC
  - COVID-19 Case Surveillance Public Use Data | Data | Centers for Disease Control and Prevention
- Conglomeration of datasets from the UN, European CDPC, Johns Hopkins, World Bank and national government reports
  - Our World in Data - Coronavirus Source Data
- World Economic Outlook Database (Bi - Annual)
  - IMF - World Economic Outlook Database
- A large-scale COVID-19 Twitter chatter dataset for open scientific research and international collaboration from Panacea Lab - Georgia State University
  - Zenodo link
- API call formatting and updated statistics for live updating application
  - Disease.sh - provides formatting for proper API calls
  - worldometers - provides statistics updated every 10 minutes

CDC COVID-19 Reporting

- States were told to stop reporting to the CDC after Oct. 6 2020 and instead to report directly to US Department of Health and Human Services
  - Though the dataset continued to grow after this date
  - Possibly some states ignored this order
  - Also possible that the CDC simply began getting their information form HHS as opposed to directly from the states
- CDC provides 11 columns including
  - Date Reported, Date of Positive Test, Date of Onset of Symptoms, Current Status
  - Sex, Age Range, Race/Ethnicity, Hospital Admittance (y/n), ICU Admittance (y/n)
  - Death of the Patient (y/n), Prior Medical Conditions (y/n)
- The team doing the analysis only required 3 of those columns
  - Date Reported, Age Range, and Current Status
  - We cleaned the CSV the CDC provided to include only these columns and exported to TSV
- We defaulted to TSV for all of our data due to the possibility of commas appearing in strings breaking the separation of CSV files

Our World In Data : COVID-19

- Delivered in JSON, CSV, and XML
  - We used JSON and CSV
- JSON format delivers some general statistics regarding each nation, in addition to daily statistics from 01/01/2020 to the current day
  - The structure of the JSON made it extremely complicated to attempt parsing via Spark's built in JSON parsing
  - We used the JSHON command line tool to extract all of the general statistics for each country and write them to a tab separated file (TSV)
- CSV format delivers only the daily statistics from 01/01/2020 to present day
  - The dataset was highly denormalized due to the fact that the dataset reaches back farther than the start of the pandemic
  - We used a k8 cron job to get the most recent update to the dataset every night as a CSV
  - This file is then normalized by inserting NULL into empty columns then exported into TSV

- Country General Statistics
  - CONTINENT, COUNTRY, POPULATION, POP_DENSITY
  - MEDIAN_AGE, OLDER65, OLDER70, EXTREME_POVERTY,
  - GDP_PER_CAPITA, CARDIOVASCULAR_DEATH_RATE,
  - DIABETES_PREVALENCE, FEMALE_SMOKERS, MALE_SMOKERS
  - HANDWASHING_FACILITIES, HOSPITAL_BEDS_PER_THOUSAND
  - LIFE_EXPECTANCY, HUMAN_DEVELOPMENT_INDEX
- Country Daily Stats
  - ISO_CODE, CONTINENT, LOCATION, DATE, TOTAL_CASES
  - NEW_CASES, NEW_CASES_SMOOTHED, TOTAL_DEATHS
  - NEW_DEATHS, NEW_DEATHS_SMOOTHED, TOTAL_CASES_PER_MILLION
  - NEW_CASES_PER_MILLION, NEW_CASES_SMOOTHED_PER_MILLION
  - TOTAL_DEATHS_PER_MILLION, NEW_DEATHS_PER_MILLION
  - NEW_DEATHS_SMOOTHED_PER_MILLION, REPRODUCTION_RATE
  - ICU_PATIENTS, ICU_PATIENTS_PER_MILLION, HOSP_PATIENTS
  - HOSP_PATIENTS_PER_MILLION,
  - WEEKLY_ICU_ADMISSIONS, WEEKLY_ICU_ADMISSIONS_PER_MILLION
  - WEEKLY_HOSP_ADMISSIONS, WEEKLY_HOSP_ADMISSIONS_PER_MILLION
  - TOTAL_TESTS, NEW_TESTS, TOTAL_TESTS_PER_THOUSAND
  - NEW_TESTS_PER_THOUSAND, NEW_TESTS_SMOOTHED
  - NEW_TESTS_SMOOTHED_PER_THOUSAND POSITIVE_RATE, TESTS_PER_CASE
  - TESTS_UNITS

International Monetary Fund Data

Includes data on country's economic output and trading, such as:

- Country and Year
- Various GDP indicators
- Inflation indicators
- Import/Export indicators
- Gov't spending and borrowing
- Savings habits
- Employment data

```
name: String,
year: Int,
gdp_constPrices: Long = null,
gdp_constPrices_delta: Double = null,
gdp_currentPrices: Long = null,
gdp_currentPrices_usd: Long = null,
gdp_currentPrices_ppp: Long = null,
gdp_deflator: Int = null,
gdp_perCap_constPrices: Long = null,
gdp_perCap_constPrices_ppp: Double = null,
gdp_perCap_currentPrices: Long = null,
gdp_perCap_currentPrices_usd: Long = null,
gdp_perCap_currentPrices_ppp: Double = null,
output_gap_pGDP: Double = null,
gdp_ppp_frac_of_total_world: Double = null,
implied_ppp: Double = null,
total_investment: Double = null,
gross_national_savings: Double = null,
inflation_avgConsumerPrices: Double = null,
inflation_avgConsumerPrices_delta: Double = null,
inflation_eopConsumerPrices: Double = null,
inflation_eopConsumerPrices_delta: Double = null,
six_month_LIBOR: Double = null,
vol_imports_goods_and_services_delta: Double = null,
vol_imports_goods_delta: Double = null,
vol_exports_goods_and_services_delta: Double = null,
vol_exports_goods_delta: Double = null,
unemployment_rate: Double = null,
employed_persons: Long = null,
population: Long = null,
government_revenue_currency: Long = null,
government_revenue_percent: Double = null,
government_total_expenditure_currency: Long = null,
government_total_expenditure_percent: Double = null,
government_net_lb_currency: Long = null,
government_net_lb_percent: Double = null,
government_structural_balance_currency: Long = null,
government_structural_balance_percent_pGDP: Double = null,
government_primary_net_lb_currency: Long = null,
government_primary_net_lb_percent: Double = null,
government_net_debt_currency: Long = null,
government_net_debt_percent: Double = null,
government_gross_debt_currency: Long = null,
government_gross_debt_percent: Double = null,
gdp_of_fiscal_year: Long = null,
current_account_balance_usd: Long = null,
current_account_balance_percentGDP: Double = null
```

Dictionaries and Case Classes

Dictionaries and Lists:

- Bordering Countries
- Landlocked Countries
- Country Codes
- Reverse Country Codes
- Regions
- Development Rankings

Including methods for:

- Country Code Lookup
- Bordering Countries
- Is Landlocked
- Is in Region

Case Classes:

- Region
  - ↳ Name, Agg. Population, Avg. Median Age,...
- Country
  - ↳ Name, Population, Median Age,...
- Case Data
  - ↳ Country, Date, Total Cases, Total Deaths,...
- Economics Data
  - ↳ Country, Year, GDP (various), Export/Import,...
- Tweet
  - ↳ Timestamp, Id, Text

Twitter's Data

- ~200 million dehydrated tweets / ~20 million hydrated tweets
  - Panacea full-data set (Zenodo) > Cleaned our data to just tweet Ids > created a sample dataset with every 13th tweet > hydrated tweets > uploaded hydrated dataset to s3
- Tweets are received as JSON
  - Contains large amount of complex data which is primarily irrelevant to our analyses
  - Extra data includes source, reply status, user, quote status, and others
  - Timestamp, ID, Text, and Country Code are stored in Tweet Case Class
    - ↪️ Country Code is obtained from the optional TwitterPlace object, which is not handled otherwise
  - Tweets objects are able to return hashtags via RegEx with the getHashtags method

Question 1

Blue Team: Liam Hood, D’Ante Jolly, Sean Tidd, Nahshon Williams

Objective:

- Which regions handled COVID-19 the best using the metrics of percentage change in GDP and COVID-19 infection rate?

Additional exploration:

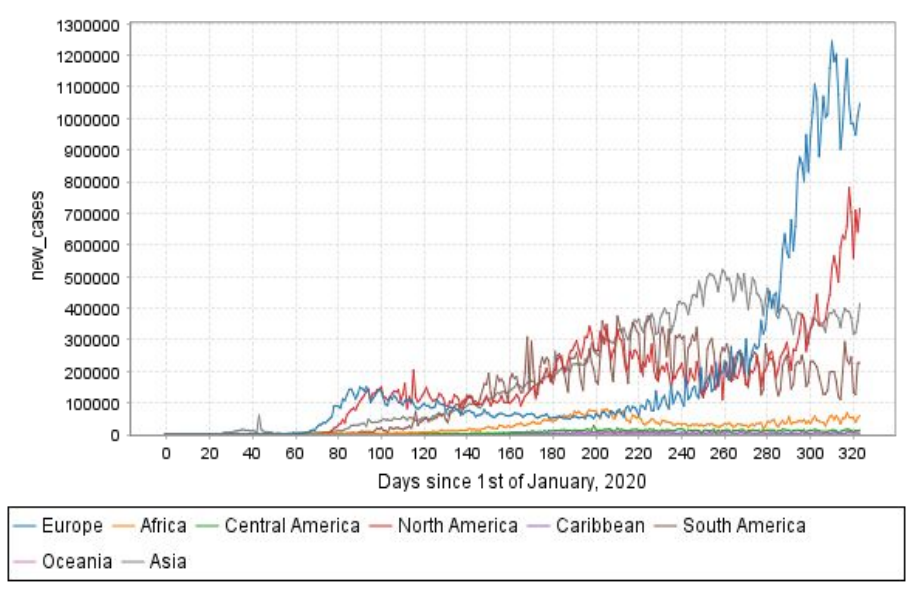
- We changed to using per capita infection rate
- We added the total case count as an additional metric

Question 1

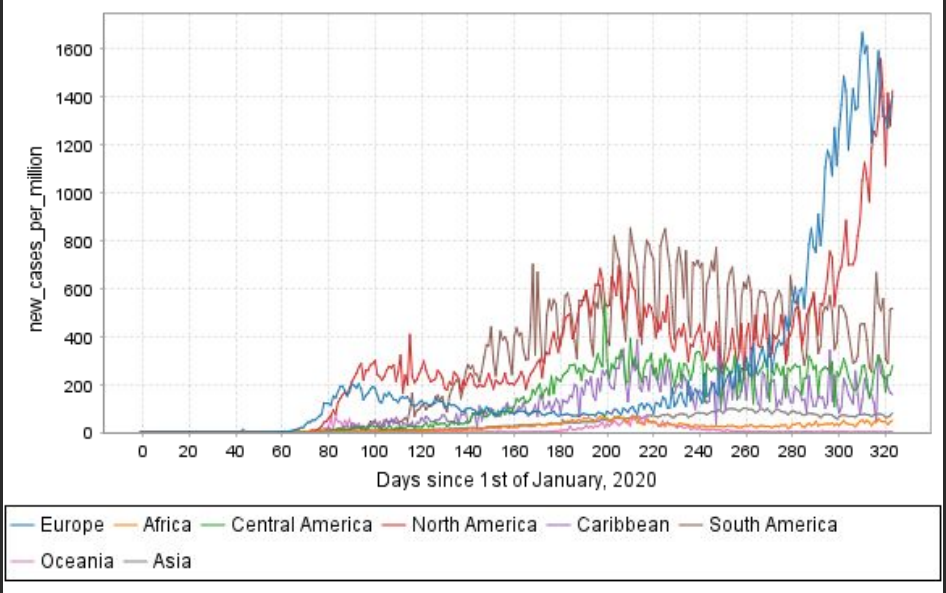
Part B

Regions compared by daily infection rate

- The totals of the graph are new cases per day which gives us the infection rate.



New Cases Daily by Region



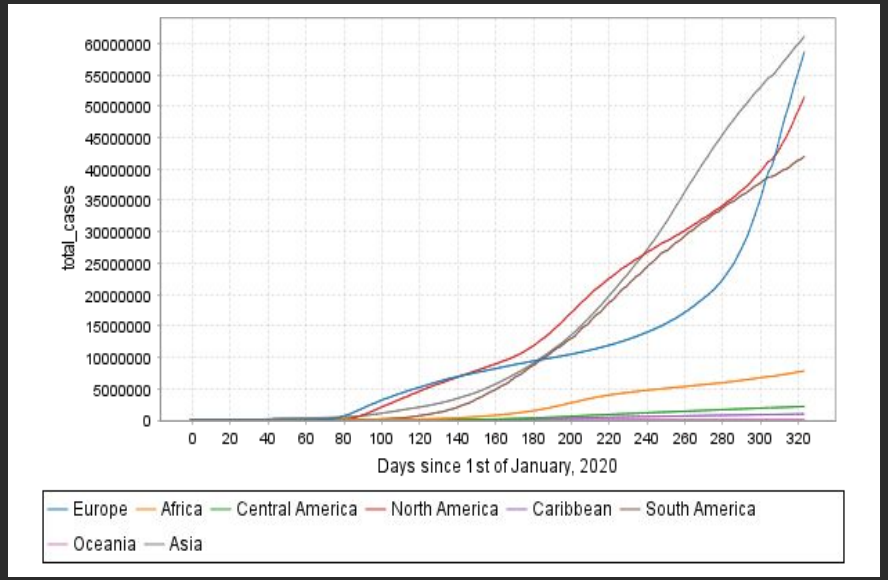
New Cases Daily by Region Normalized

Question 1

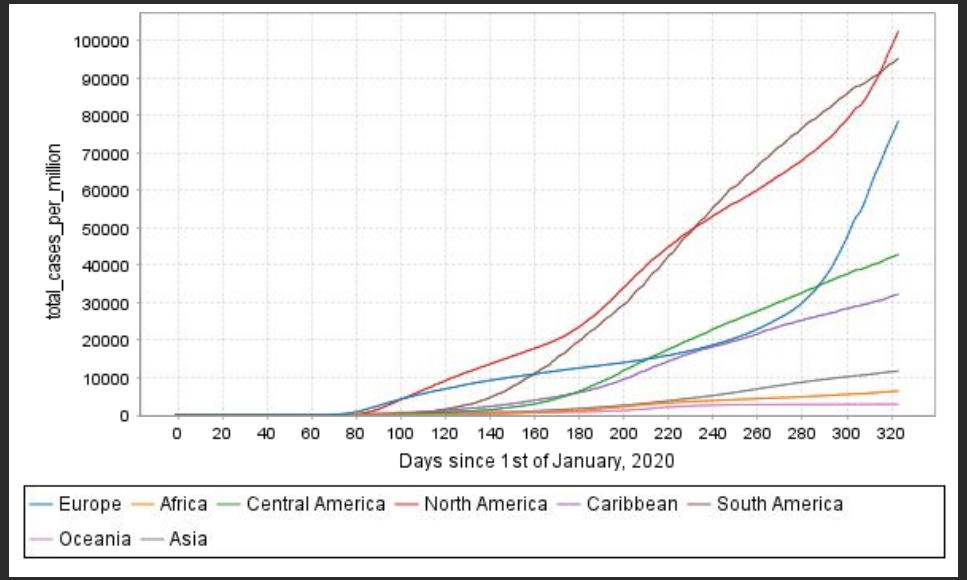
Part B

Regions compared by total cases starting from day 1 of January 2020

- Data on these graphs is total overall cases. The graph on the right is normalized by one million.



Total Cases per Region

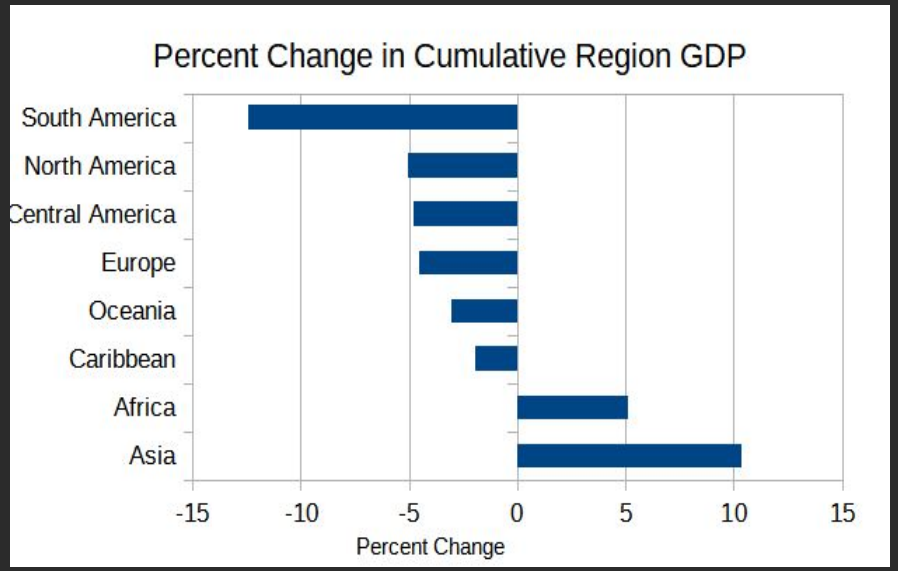


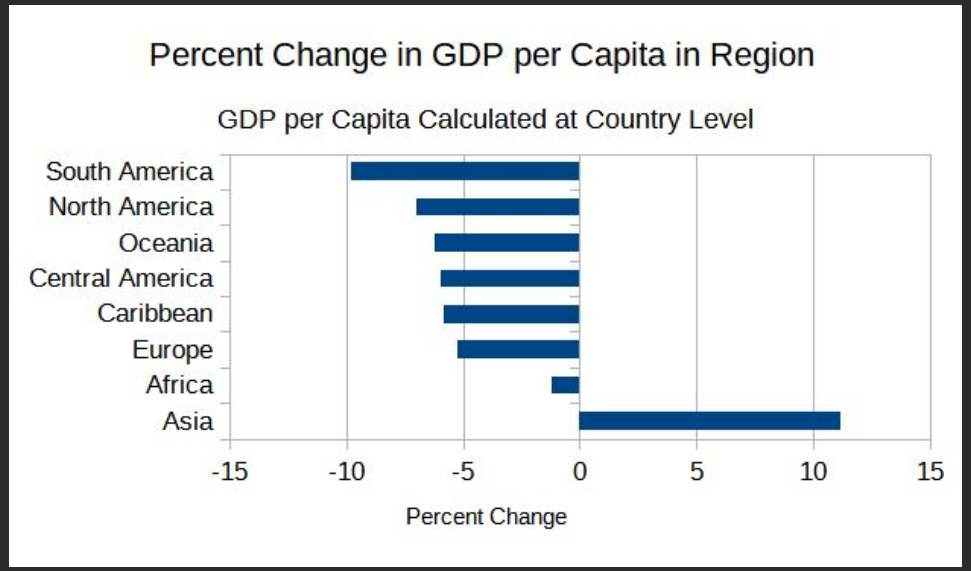
Total Cases per Region Normalized

Question 1

Regions compared by percent change in GDP from 2019 to 2020

Part B





Question 2

Orange Team: Chris Chee, Roger Griffin, Jordan Juel, Edward Reed

Objectives:

- Find the top 5 pairs of countries that share a land border and have the highest discrepancy in COVID-19 infection rate per capita.
- Find the top 5 landlocked countries that have the highest COVID-19 infection rate per capita.

Additional exploration:

- Find the COVID-19 infection rate per capita of island countries
- Compare COVID-19 infection rate per capita based on the Human Development Index

Question 2

Find the top 5 pairs of countries that share a land border and have the highest discrepancy in COVID-19 infection rate per capita.

countries_general_stats.tsv

COUNTRY	POPULATION
Afghanistan	38928341
Albania	2877800
Algeria	43851043
Andorra	77265
Angola	32866268
Anguilla	15002

Resulting population table

JOIN

daily_stats.tsv

COUNTRY	TOTAL CASES
Afghanistan	45017
Albania	33556
Algeria	75867
Andorra	6304
Angola	14493
Anguilla	4

Resulting infections table

Question 2

Find the top 5 pairs of countries that share a land border and have the highest discrepancy in COVID-19 infection rate per capita.

COUNTRY	TOTAL CASES	POPULATION	INFECTION RATE PER CAPITA
Afghanistan	45017	38928341	0.11564068450797839
Albania	33556	2877800	1.1660296059489887
Algeria	75867	43851043	0.17301070809193753
Andorra	6304	77265	8.158933540412864
Angola	14493	32866268	0.04409688377153135
Anguilla	4	15002	0.026663111585121985

Resulting table from join

```
val borders_dictionary = Map[String, List[String]](
    elems = "Andorra" -> List(
        "France",
        "Spain"
    ),
    "United Arab Emirates" -> List(
        "Oman",
        "Saudi Arabia"
    ),
    "Afghanistan" -> List( //Afgahnastan
        "China",
        "Iran",
        "Pakistan",
        "Tajikistan", //tajikastan
        "Turkmenistan",
        "Uzbekistan"
    ),
    "Antigua and Barbuda" -> List(),
    "Anguilla" -> List(),
```

Contains border information on all countries

Question 2

Part A

Find the top 5 pairs of countries that share a land border and have the highest discrepancy in COVID-19 infection rate per capita.

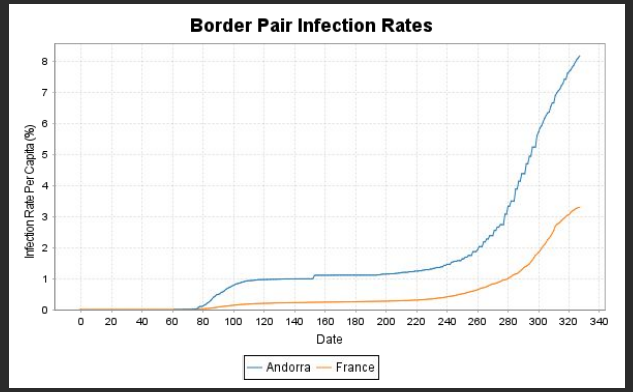
country_name	country_infection_rate	border_country	country_border_infection_rate_per_capita	delta
Andorra	8.159	France	3.286	4.873
Andorra	8.159	Spain	3.385	4.774
Luxembourg	4.942	Latvia	0.702	4.240
Montenegro	5.034	Albania	1.166	3.868
Israel	3.816	Syria	0.042	3.774

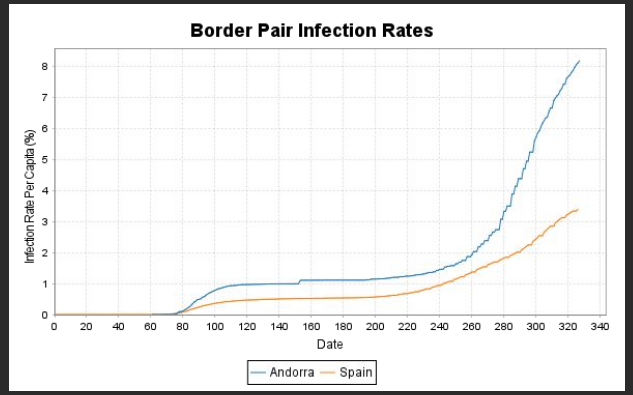
Final Result

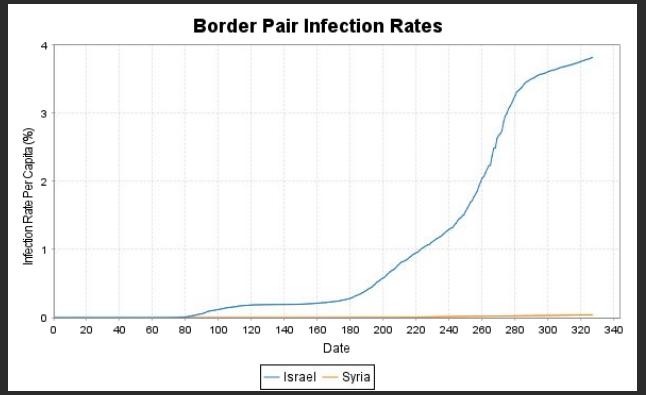
Question 2

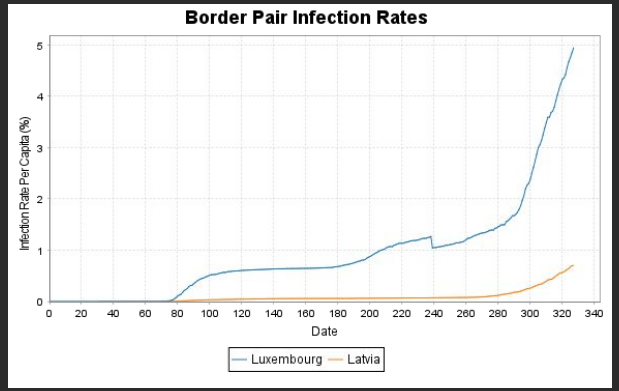
Part A

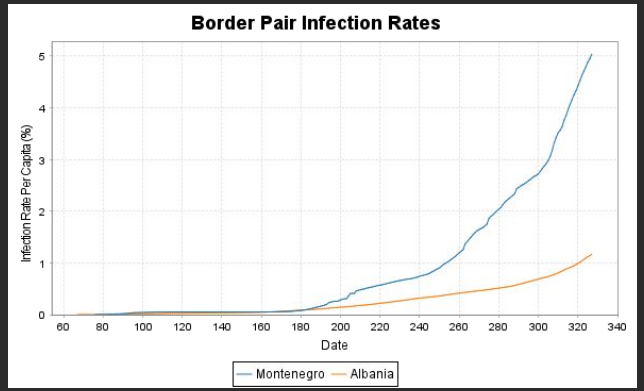
Find the top 5 pairs of countries that share a land border and have the highest discrepancy in COVID-19 infection rate per capita.











Question 2

Part B

Find the top 5 landlocked (and island) countries that have the highest COVID-19 infection rate per capita.

Highest Infection Rate in Land Locked Countries

country_name	infection_rate_per_capita(%)
Andorra	8.159
Luxembourg	4.942
Czech Republic	4.638
Armenia	4.303
San Marino	4.208

Highest Infection Rate in Water Locked Countries

country_name	infection_rate_per_capita(%)
Bahrain	5.047
French Polynesia	4.620
Aruba	4.437
Guam	3.912
Puerto Rico	3.008

Question 2

Part C

Compare COVID-19 infection rate per capita based on the Human Development Index

Human Development Index (HDI) is a measure of achievement for a country based on three dimensions of human development.

- Life expectancy at birth.
- How many years of education a citizen is expected to receive.
- Gross national income per capita.

Question 2

Part C

Compare COVID-19 infection rate per capita based on the Human Development Index

Largest infection rates with the highest ranking countries by HDI (Human Development Index)

country_name	infection_rate_per_capita
Andorra	8.159
Bahrain	5.047
Montenegro	5.034
Luxembourg	4.942
Belgium	4.828

Highest infection rate with average ranking countries

country_name	infection_rate_per_capita
Armenia	4.303
Panama	3.608
Peru	2.883
Brazil	2.864
Georgia	2.803

Largest infection rate with the lowest ranking countries

country_name	infection_rate_per_capita
Honduras	1.059
Nepal	0.763
India	0.665
Djibouti	0.574
Namibia	0.547

Question 3

Red Team: Ernie Chu, Kevin Conlin, John Rice, Syed Rizvi

Objective:

- Provide live updates by Region of current relevant totals from COVID-19 data

Relevant Totals:

- Total Cases
- New Daily Cases
- Total Deaths
- New Daily Deaths
- Total Recoveries
- New Daily Recoveries

Regions:

- Africa
- Asia
- Caribbean
- Central America
- Europe
- North America
- Oceania
- South America

Question 3

Provide live updates of relevant COVID-19 statistics

Using the disease.sh API, newly updated data can be pulled in every 10 minutes in JSON format to provide the most up-to-date data from Worldometer.

These regional JSONs are loaded into DF to create tables for each individual regions relevant statistics: Total, New, and Percent increase for Cases, Fatalities, and Recoveries respectively.

Through the use of Spark SQL, these DataFrames are unioned and the appropriate calculations are made to provide a final table representing the stats from each region as well as the total aggregated stats.

Each table is time stamped with the date and time it was created in UTC.

Question 3

Provide live updates of relevant COVID-19 statistics

Output Table

Last Updated: 2020-12-07 17:16:44

Region	Total Cases	Today's Cases	Cases Percent Change	Total Deaths	Today's Deaths	Death Percent Change	Total Recoveries	Today's Recoveries	Recoveries Percent Change
Europe	18447384	126309	0.68	425466	2448	0.58	7745589	100322	1.3
Asia	17402954	97149	0.56	299972	1289	0.43	15390413	76560	0.5
Caribbean	206515	755	0.37	3369	1	0.03	155786	680	0.44
North America	16788043	43637	0.26	411551	513	0.12	10071953	26379	0.26
Africa	2253043	1699	0.08	53665	24	0.04	1917171	1784	0.09
Central America	611839	424	0.07	13684	30	0.22	457431	527	0.12
South America	11541199	3584	0.03	332489	82	0.02	10257675	1681	0.02
Oceania	45935	8	0.02	1021	0	0.0	33157	4	0.01
Total	67296912	273565	0.26	1541217	4387	0.18	46029175	207937	0.34

Log output containing the most up-to-date table can be found with the command:
kubectl logs q3-bf10e1762a7ec2af-driver | tail -n 17

Question 3 - Analysis

Provide live updates of relevant COVID-19 statistics

After running the application for 1 week to monitor trend patterns in the data, several patterns became apparent:

- North America and Europe consistently finished at the top of the chart for rate of increase in new cases each day.
- Oceania remained at the bottom of the chart in terms of both percentage increase in new cases, as well as pure magnitude of new cases.
- North America and Europe provided the most frequent updates to their statistics, with Asia, Africa, and the Caribbean providing updates slightly less frequently, and South America providing the least frequent updates. (No pattern could be drawn with regard to Oceania's reporting frequencies, as their numbers changed very little throughout a given day).
- Very little reporting occurs from 2AM to 6AM UTC with the pace rapidly increasing between 6AM and 9AM before hitting a relatively stable reporting rate that remains for the rest of the day.

Question 4

Green Team: Brandon Linton, Zach Minnich, Victor Pullas, Quan Vu

Objectives:

- Is the trend of global COVID-19 discussion going up or down?
- Do spikes in infection rates of the 5-30 age range affect the volume of discussion?

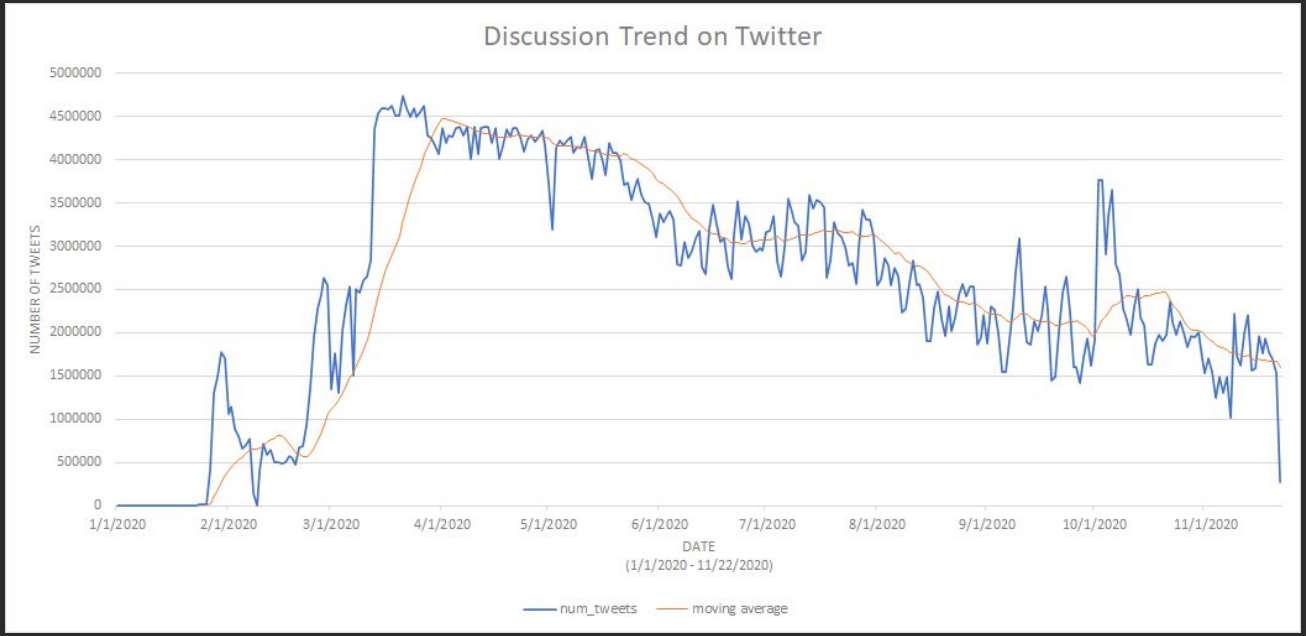
Additional exploration:

- Trend of discussion since peak
- Trend of discussion over the last reported month
- Trend of discussion over the last reported week

Question 4

Part A

Is the trend of the global COVID-19 discussion going up or down?

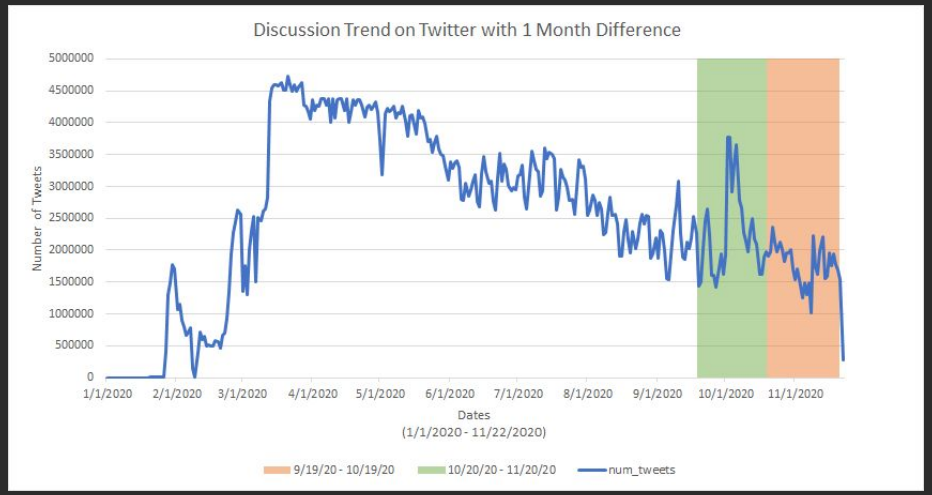


Question 4

Part A

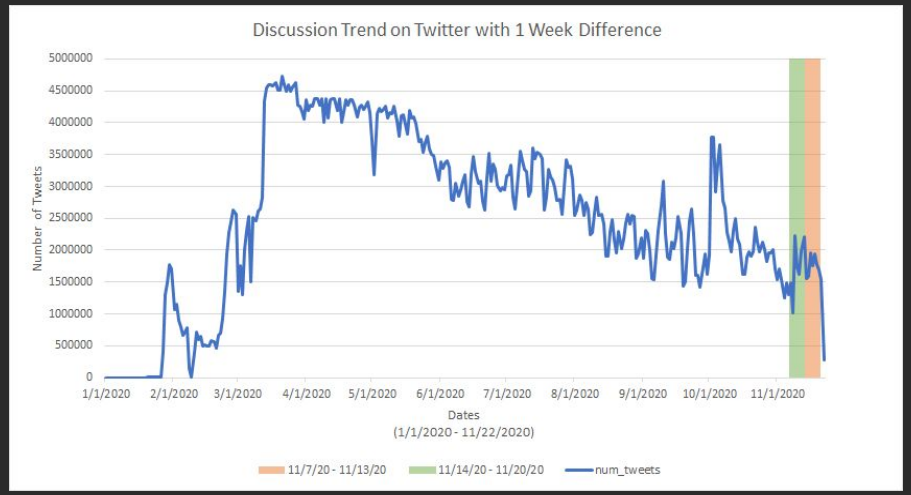
Is the trend of the global COVID-19 discussion going up or down?

2 Month Comparison



-20% last 60 days

2 Week Comparison

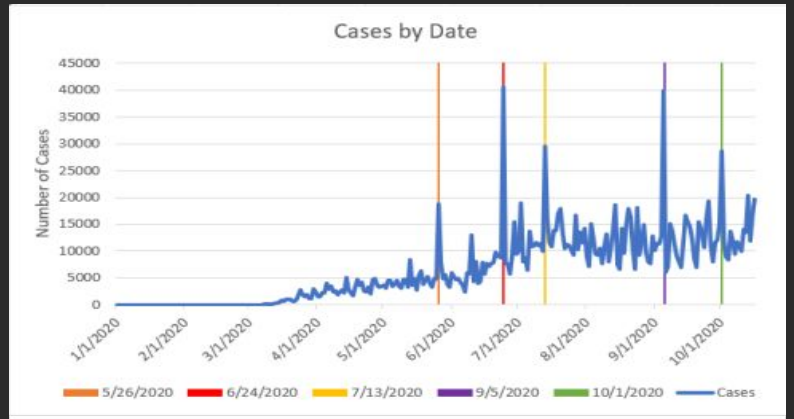


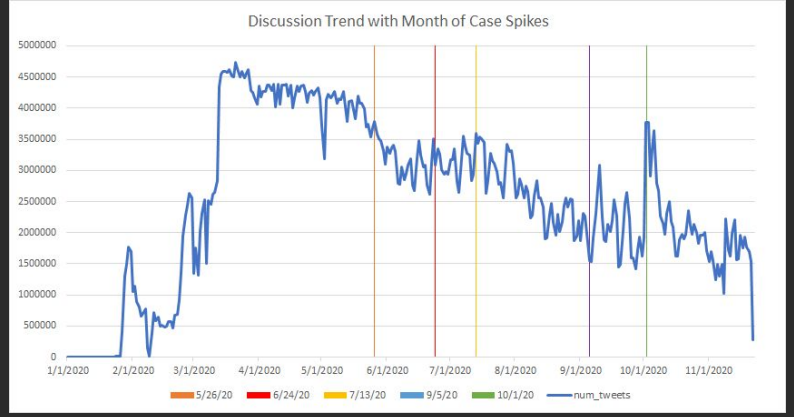
>-1% last 14 days

Question 4

Part B

Do spikes in infection rates of the 5-30 age range affect the volume of discussion?

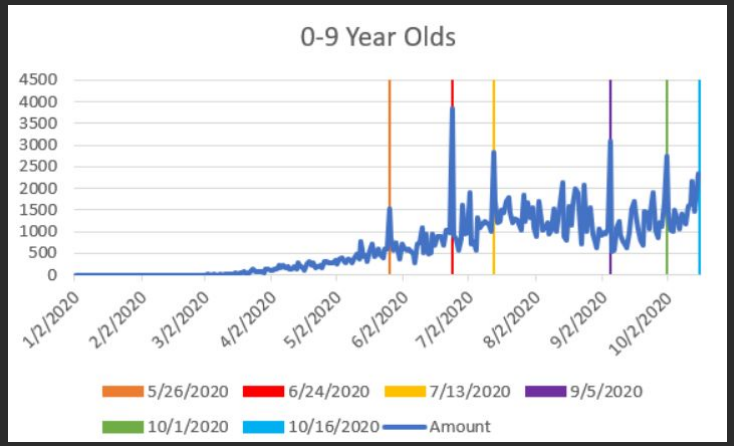


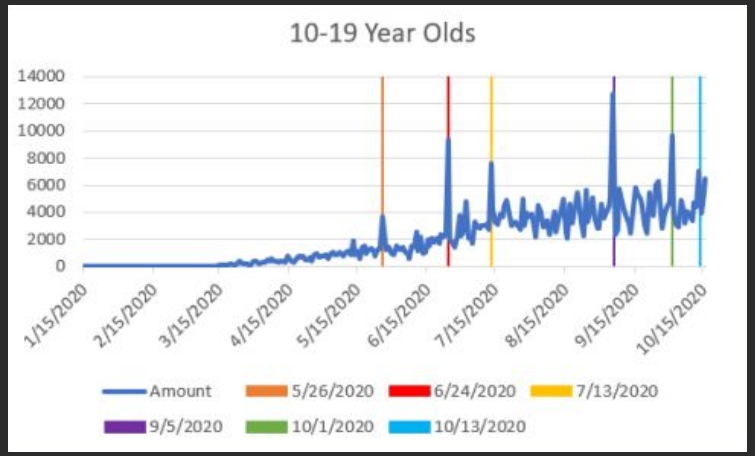


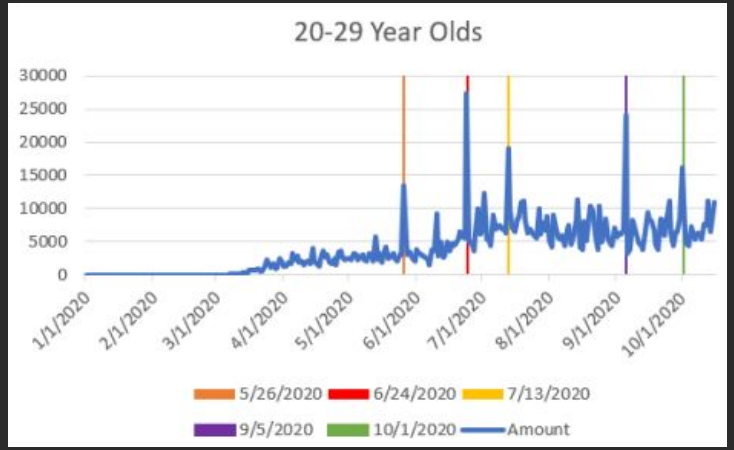
Date	Amount of Cases
6/24/2020	40557
9/5/2020	39832
7/13/2020	29507
10/1/2020	28542
10/13/2020	20385
10/16/2020	19658
9/25/2020	19345
7/2/2020	18943
5/26/2020	18686
8/14/2020	18549
8/24/2020	18023
7/20/2020	17881
8/20/2020	17816
7/19/2020	16967
9/15/2020	16747
7/27/2020	16739
10/15/2020	16663
8/21/2020	16384
9/24/2020	15985
7/14/2020	15891

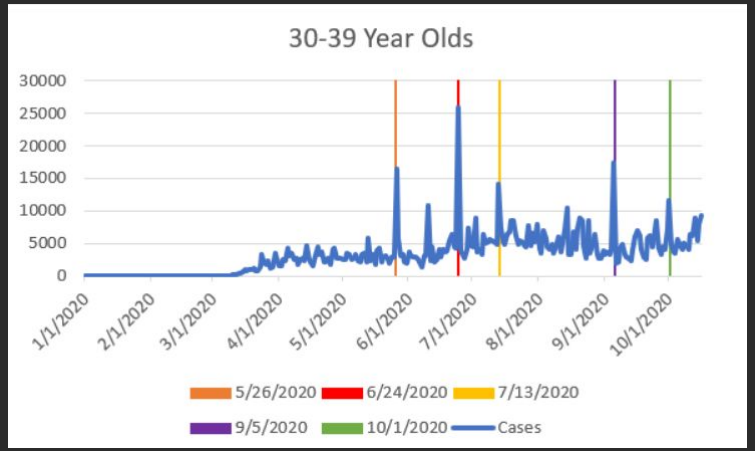
Question 4

Part B









Question 5

Green Team: Brandon Linton, Zach Minnich, Victor Pullas, Quan Vu

Objective:

- When was COVID-19 being discussed the most?

Additional exploration:

- Finding the peaks by **months** and **days** starting from January 1, 2020 to the end of November 2020.

- Finding the peak hour of the day that has highest COVID-19 discussion.

Question 5

When was COVID being discussed the most?

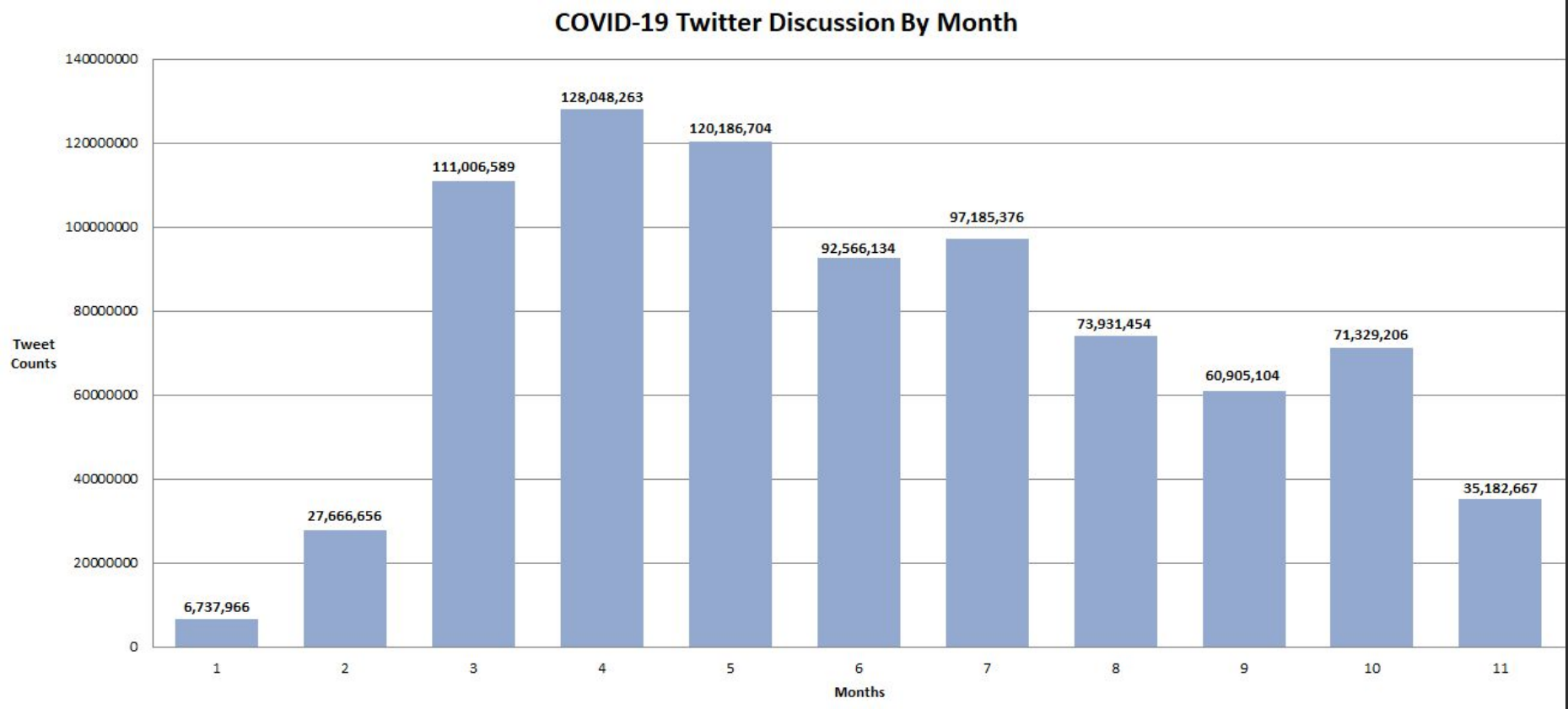
months	tweets
04	128048263
05	120186704
03	111006589
07	97185376
06	92566134
08	73931454
10	71329206
09	60905104
11	35182667
02	27666656
01	6737966

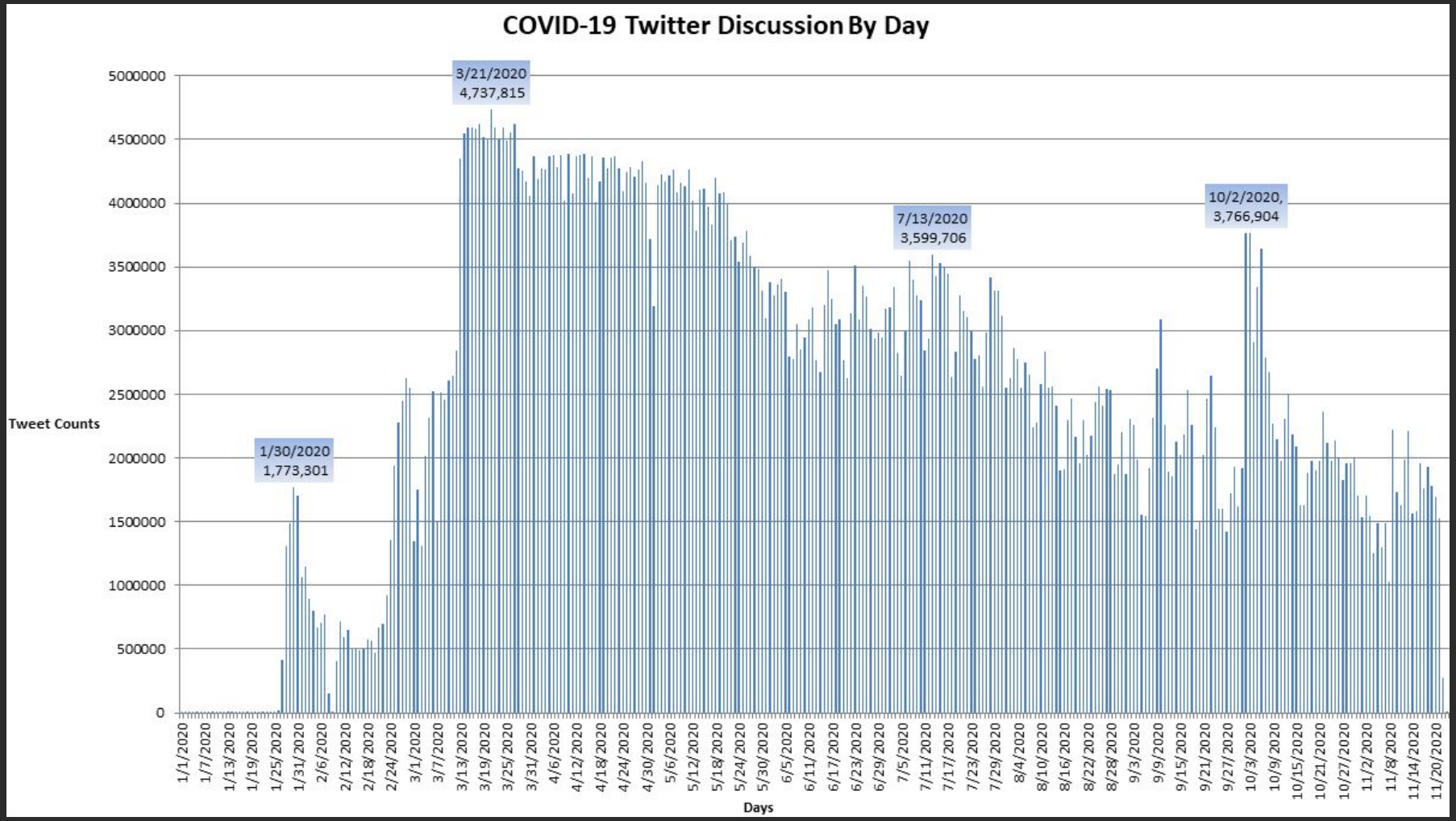
days	tweets
2020-03-21	4737815
2020-03-18	4625664
2020-03-27	4621539
2020-03-15	4596412
2020-03-22	4595815
2020-03-16	4595739
2020-03-24	4589240
2020-03-17	4581466
2020-03-26	4554228
2020-03-14	4544016
2020-03-19	4513770
2020-03-20	4509157
2020-03-23	4501584
2020-03-25	4492680
2020-04-10	4385107
2020-04-14	4381637
2020-04-08	4380089
2020-04-06	4376721
2020-04-13	4375774
2020-04-16	4370988

only showing top 20 rows

Question 5

When was COVID being discussed the most?




COVID-19 Twitter Discussion By Day
Tweet Counts
Days
3/21/2020
4,737,815
7/13/2020
3,599,706
10/2/2020,
3,766,904
1/30/2020
1,773,301
5000000
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
1/1/2020
1/7/2020
1/13/2020
1/19/2020
1/25/2020
1/31/2020
2/6/2020
2/12/2020
2/18/2020
2/24/2020
3/1/2020
3/7/2020
3/13/2020
3/19/2020
3/25/2020
3/31/2020
4/6/2020
4/12/2020
4/18/2020
4/24/2020
4/30/2020
5/6/2020
5/12/2020
5/18/2020
5/24/2020
5/30/2020
6/5/2020
6/11/2020
6/17/2020
6/23/2020
6/29/2020
7/5/2020
7/11/2020
7/17/2020
7/23/2020
7/29/2020
8/4/2020
8/10/2020
8/16/2020
8/22/2020
8/28/2020
9/3/2020
9/9/2020
9/15/2020
9/21/2020
9/27/2020
10/3/2020
10/9/2020
10/15/2020
10/21/2020
10/27/2020
11/2/2020
11/8/2020
11/14/2020
11/20/2020

Question 6

Red Team: Ernie Chu, Kevin Conlin, John Rice, Syed Rizvi

Objective:

- What percentage of countries have an increasing infection rate per capita?

Additional exploration:

- What are the changes in infection rate among different regions?
- What country had the **most** and **least** percentage increase in infection rate per capita?
- What country had the **most** and **least** percentage increase in fatality rate per capita?
- What country had the **most** and **least** percentage increase in recovery rate per capita?

Question 6

Part A

What are the changes in infection rate among different regions?

(today.todayCases/today.population) * 1,000,000 = today's infection rate per capita => "**T**"
(yesterday.todayCases/yesterday.population) * 1,000,000 = yesterday's infection rate per capita = "**Y**"

(T - Y) / Y * 100

Step 1	Step 2	Step 3	Step 4
For all regions, select a region.	**Calculate the change in infection rate change using the above equation.**	**Round to 2 decimal places with bround().**	**Average the result for all countries in the region and print the result.**

Region Infection Rate Change

Sunday-Monday

Regions and their change in Infection Rate

Region	Infection_Rate_Change
Caribbean	1.51
Africa	0.23
Asia	0.14
North America	0.03
Oceania	0.0
Central America	-0.1
South America	-0.11
Europe	-0.22

Monday-Tuesday

Regions and their change in Infection Rate

Region	Infection_Rate_Change
North America	0.96
Oceania	0.34
Europe	0.22
South America	0.06
Caribbean	0.04
Asia	-0.12
Africa	-0.18
Central America	-0.29

Tuesday-Wednesday

Regions and their change in Infection Rate

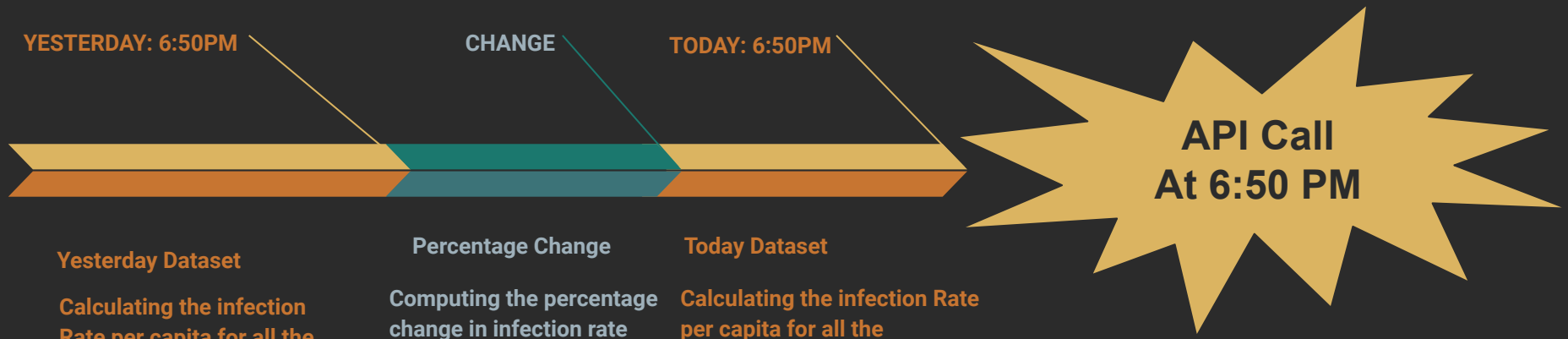
Region	Infection_Rate_Change
Europe	0.15
Africa	0.1
South America	-0.02
Asia	-0.05
Central America	-0.53
Oceania	-0.53
North America	-1.07
Caribbean	-1.18

We also wanted to ask some related questions.

1. What percentage of countries have an increasing infection rate per capita?

2. What are the countries with the **most** and **least** percentage changes in infection rate, fatality rate, and recovery rate per capita (defined as per 1,000,000).

What percentage of countries have an increasing infection rate per capita?



Yesterday Dataset

Calculating the infection Rate per capita for all the countries.

Percentage Change

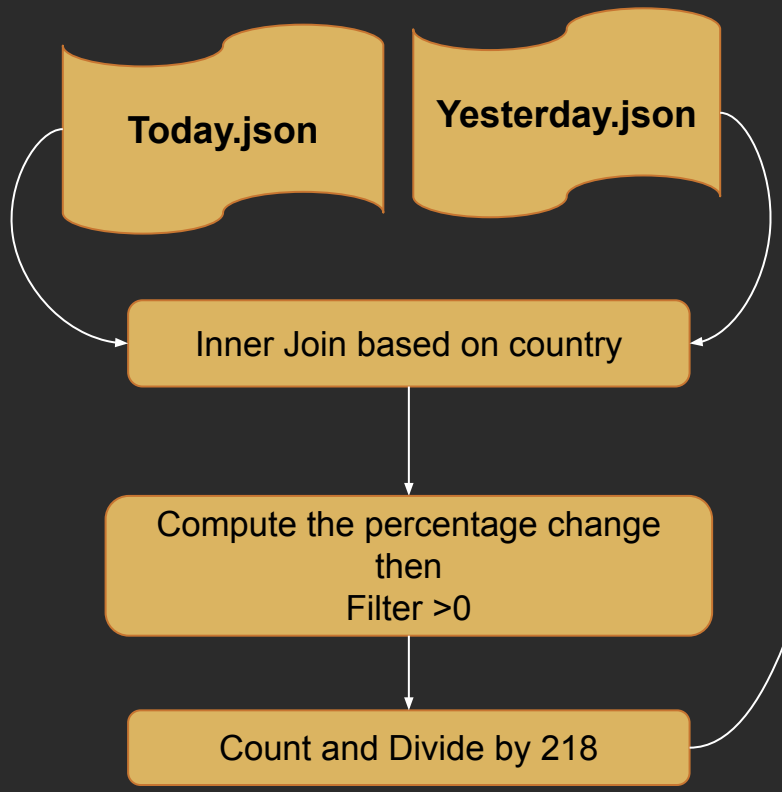
Computing the percentage change in infection rate per capita of two days.

Today Dataset

Calculating the infection Rate per capita for all the countries.

Source: https://www.worldometers.info

$$Percentage\ Change = \frac{Today\ IR - Yesterday\ IR}{Yesterday\ Cases\ Per\ Million} \times 100$$



Results

Yesterday	***Today***	***Percentage***
Sunday	Monday	39.5
Monday	Tuesday	45.4
Tuesday	Wednesday	41.7

What are the countries with the least and most percentage changes in infection rate, fatality rate, and recovery rate per capita (defined as per 1,000,000)?

Same process just:
Order By Asc/Desc

Most

Change in Infection Rate Per Capita			Change in Fatality Rate Per Capita			Change in Recovery Rate Per Capita		
Anguilla	Caribbean	28.5	Curaçao	Caribbean	14.1	Burundi	Africa	9.5
Saint Martin	Caribbean	11.7	French Polynesia	Oceania	8.8	St.Barth	Caribbean	20.9
Uganda	Africa	8.0	Uganda	Africa	6.5	Norway	Europe	25.7

Least

Change in Infection Rate Per Capita			Change in Fatality Rate Per Capita			Change in Recovery Rate Per Capita		
Channel Islands	Europe	-3.9	Cyprus	Asia	-3.3	Channel Islands	Europe	-6.9
Anguilla	Caribbean	-11.1	Curaçao	Asia	-12.4	Burundi	Africa	-8.7
Saint Martin	Caribbean	-10.4	Cyprus	Asia	-7.4	St.Barth	Caribbean	-17.3

Question 7

Purple Team: Trevor Buck, Alan Liang, Kyle Pacheco, Michael Splaver

Objectives:

- What are the hashtags used to describe COVID-19 by Region?
- What are the top 10 commonly used hashtags used alongside COVID hashtags?

Additional exploration:

- How does the most used hashtags to describe COVID-19 differ between Regions?

Question 7

Part A

What are the hashtags used to describe COVID-19 by Region?

Overview of Process

- Filter out tweets without location data (95%+)
- Map the countries to the appropriate region
- Filter to only the region we're targeting
- Group the hashtags together and count them
- Sort by most counted hashtags to get top 10

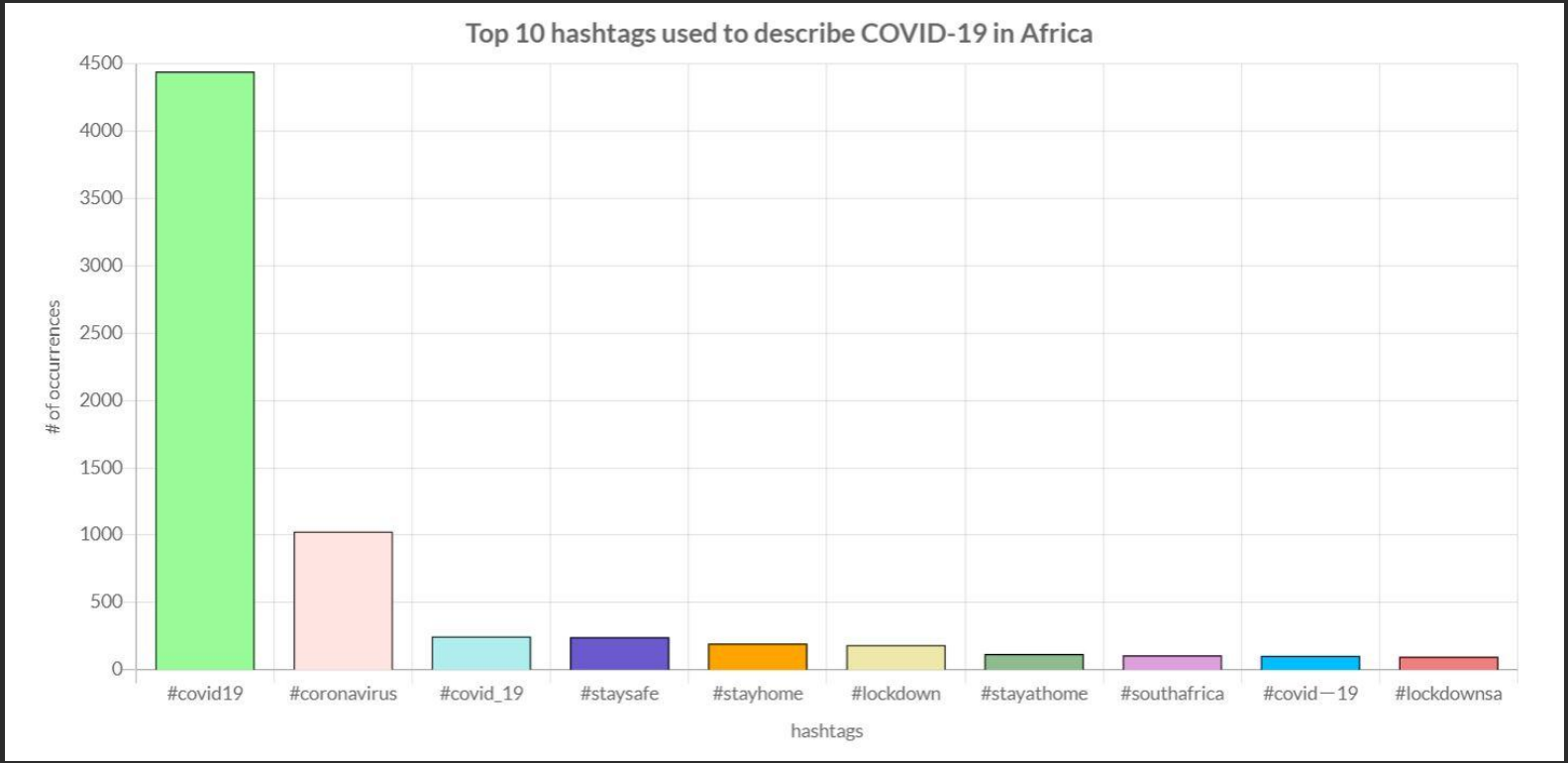
Regions used:

1. Africa
2. Asia
3. The Caribbean
4. Central America
5. Europe
6. North America
7. South America
8. Oceania

Question 7

Part A

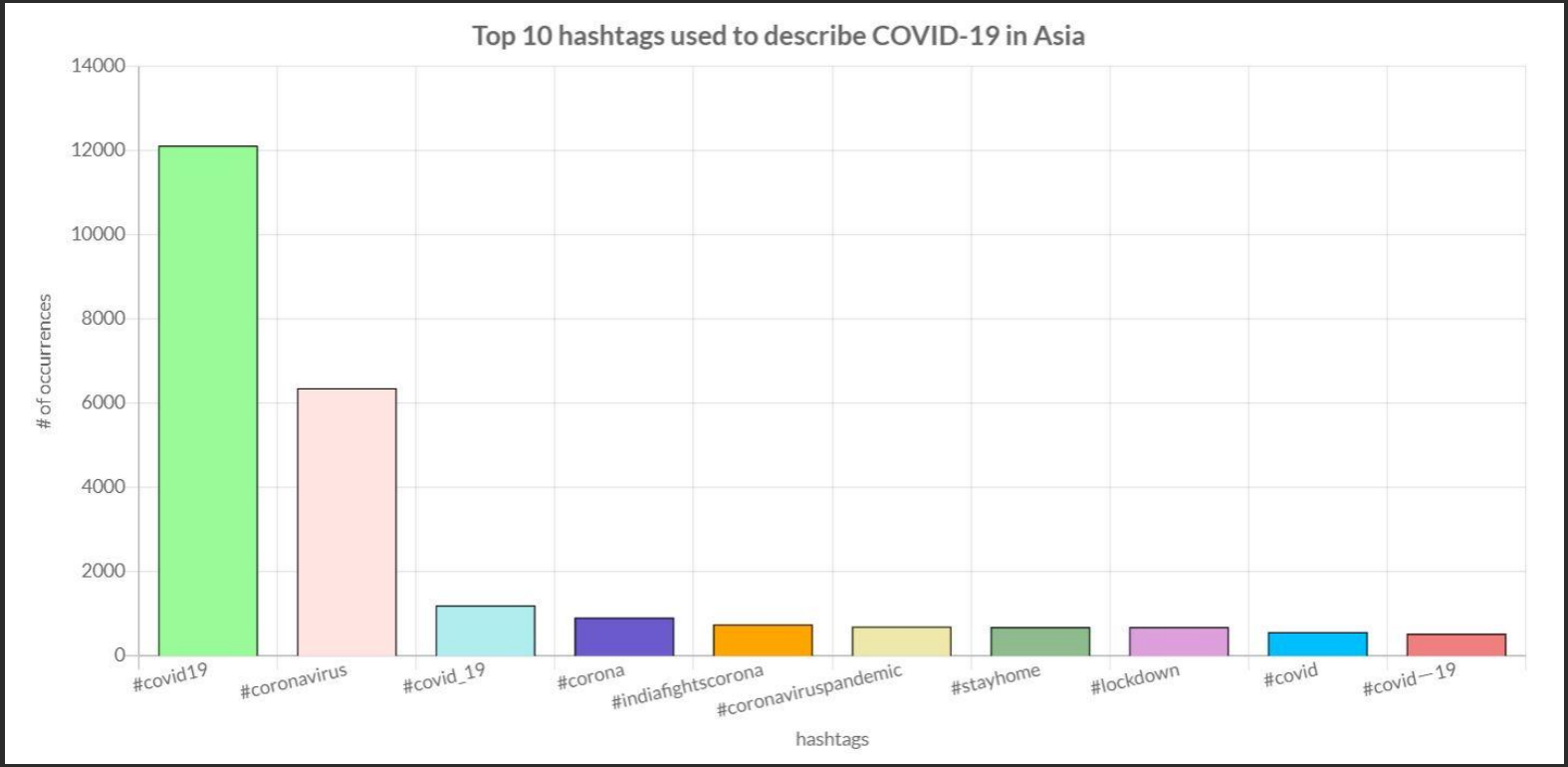
What are the hashtags used to describe COVID-19 by Region?



Question 7

Part A

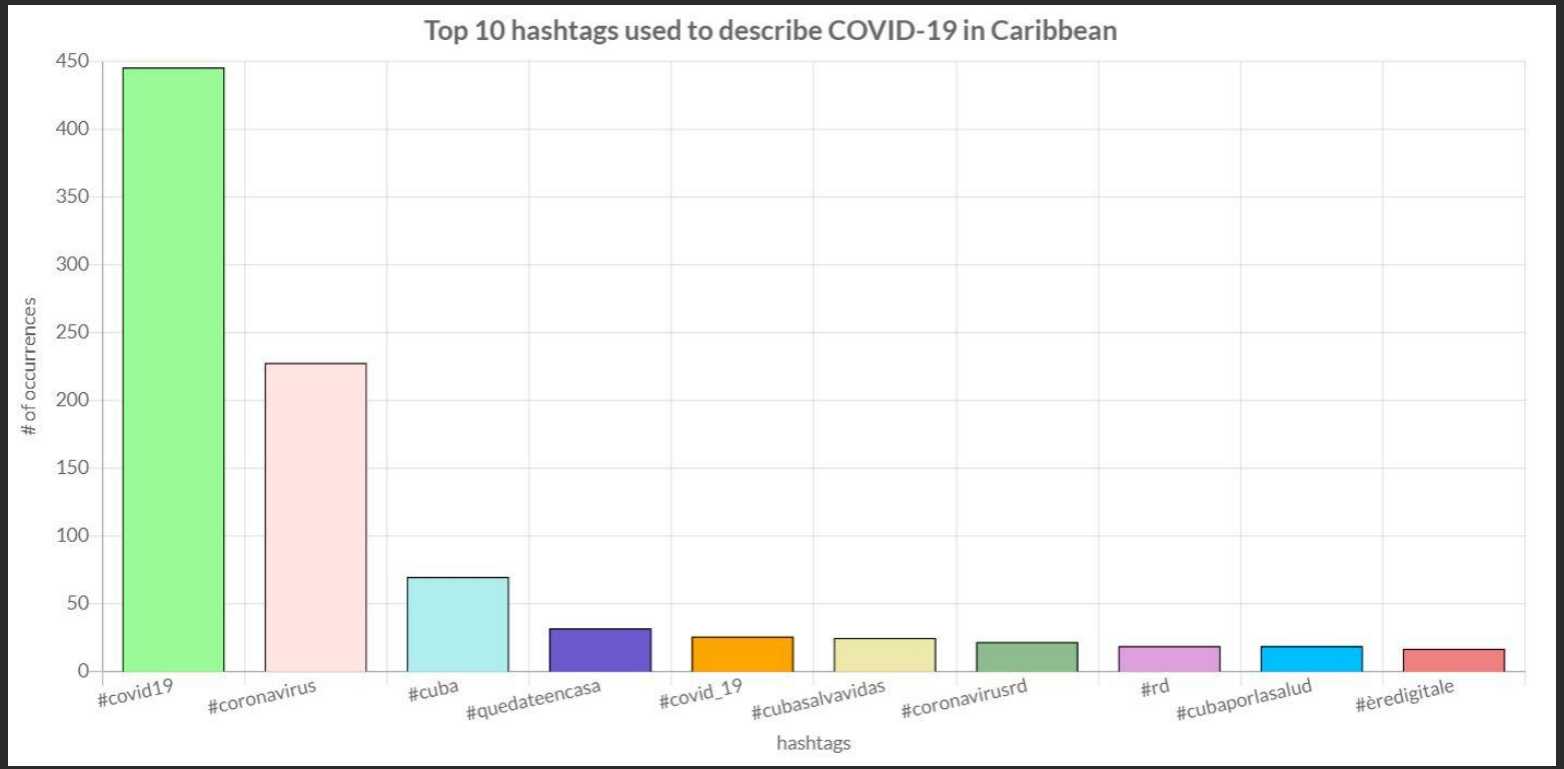
What are the hashtags used to describe COVID-19 by Region?



Question 7

Part A

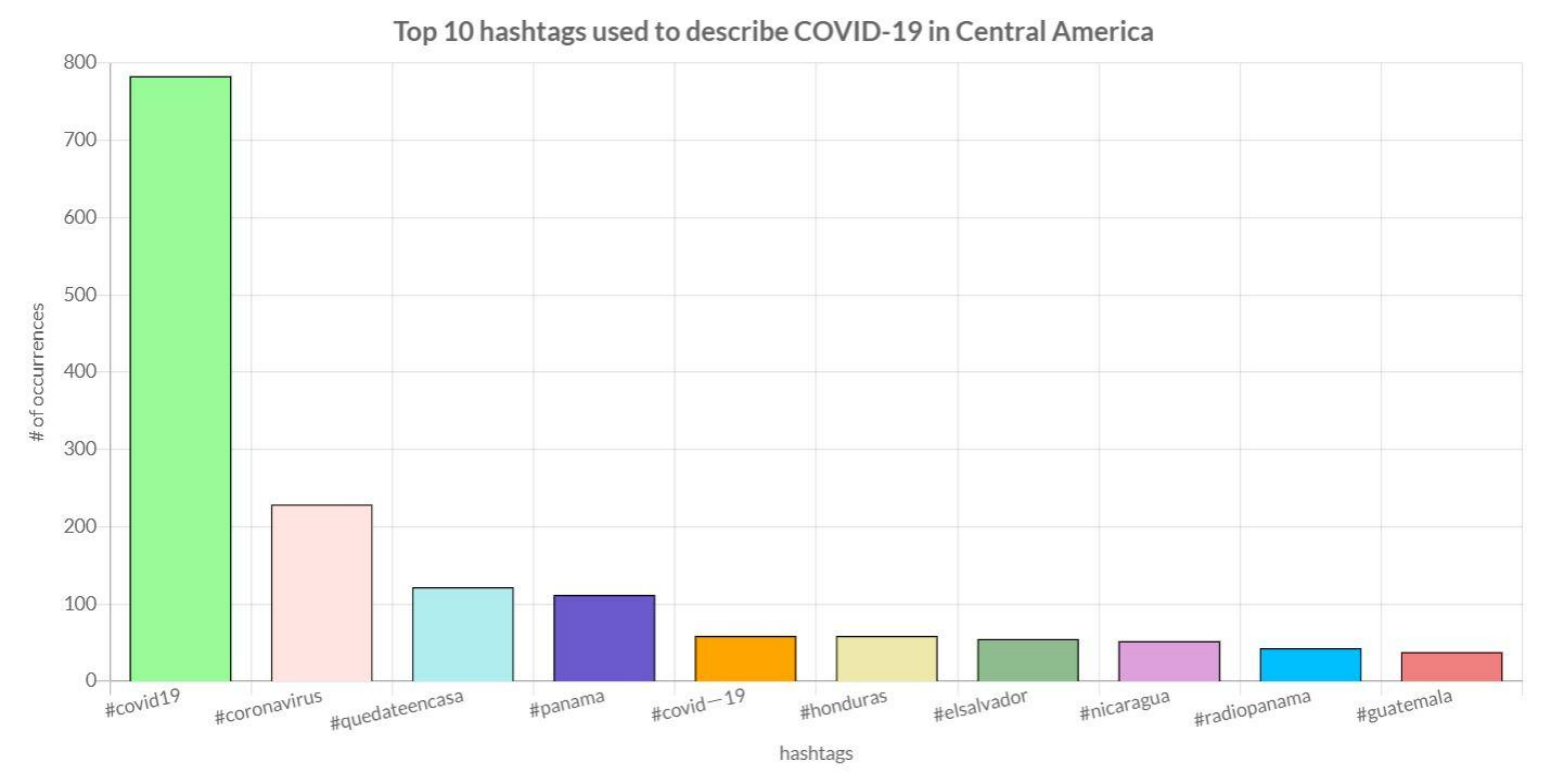
What are the hashtags used to describe COVID-19 by Region?



Question 7

Part A

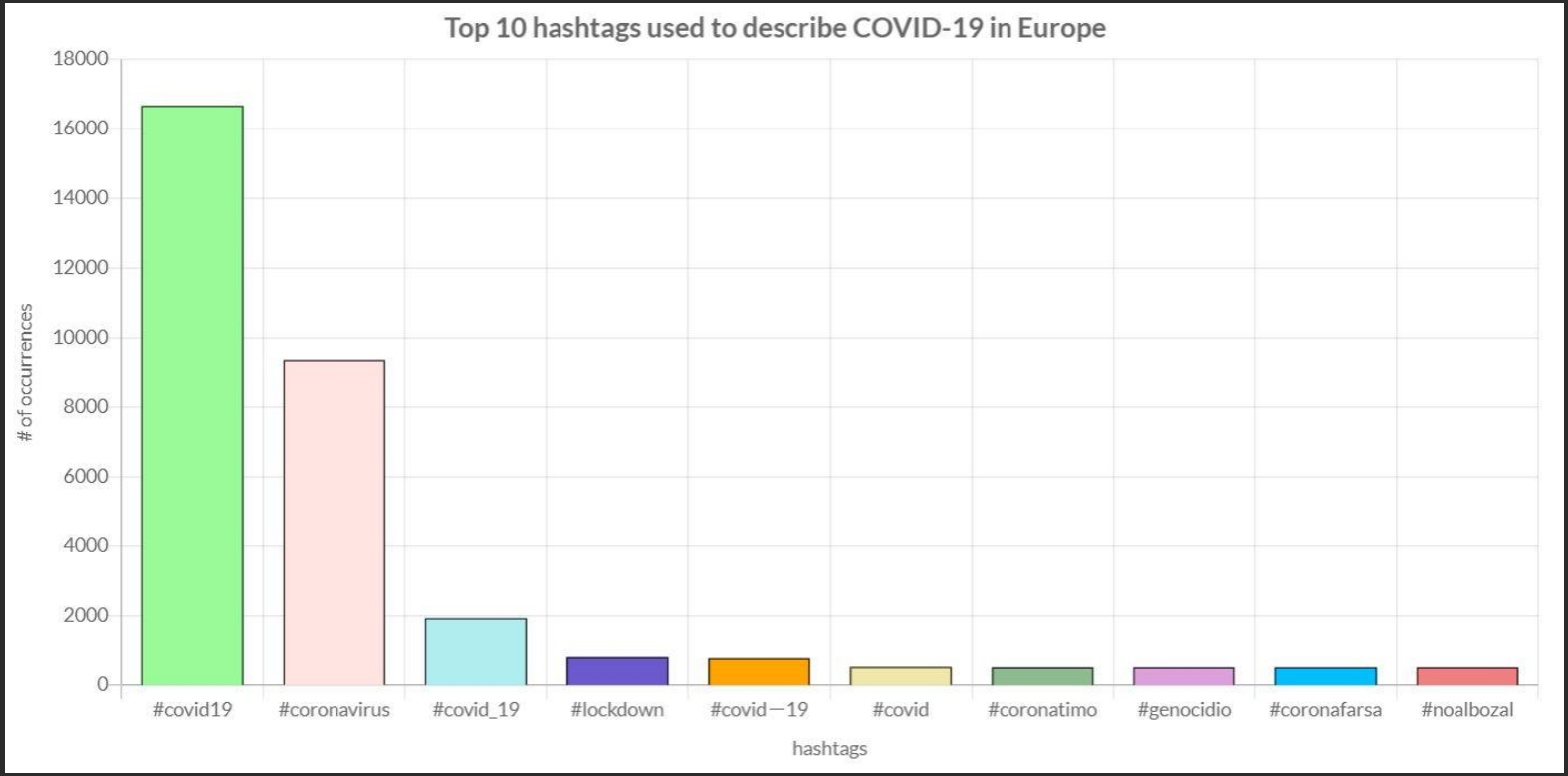
What are the hashtags used to describe COVID-19 by Region?



Question 7

Part A

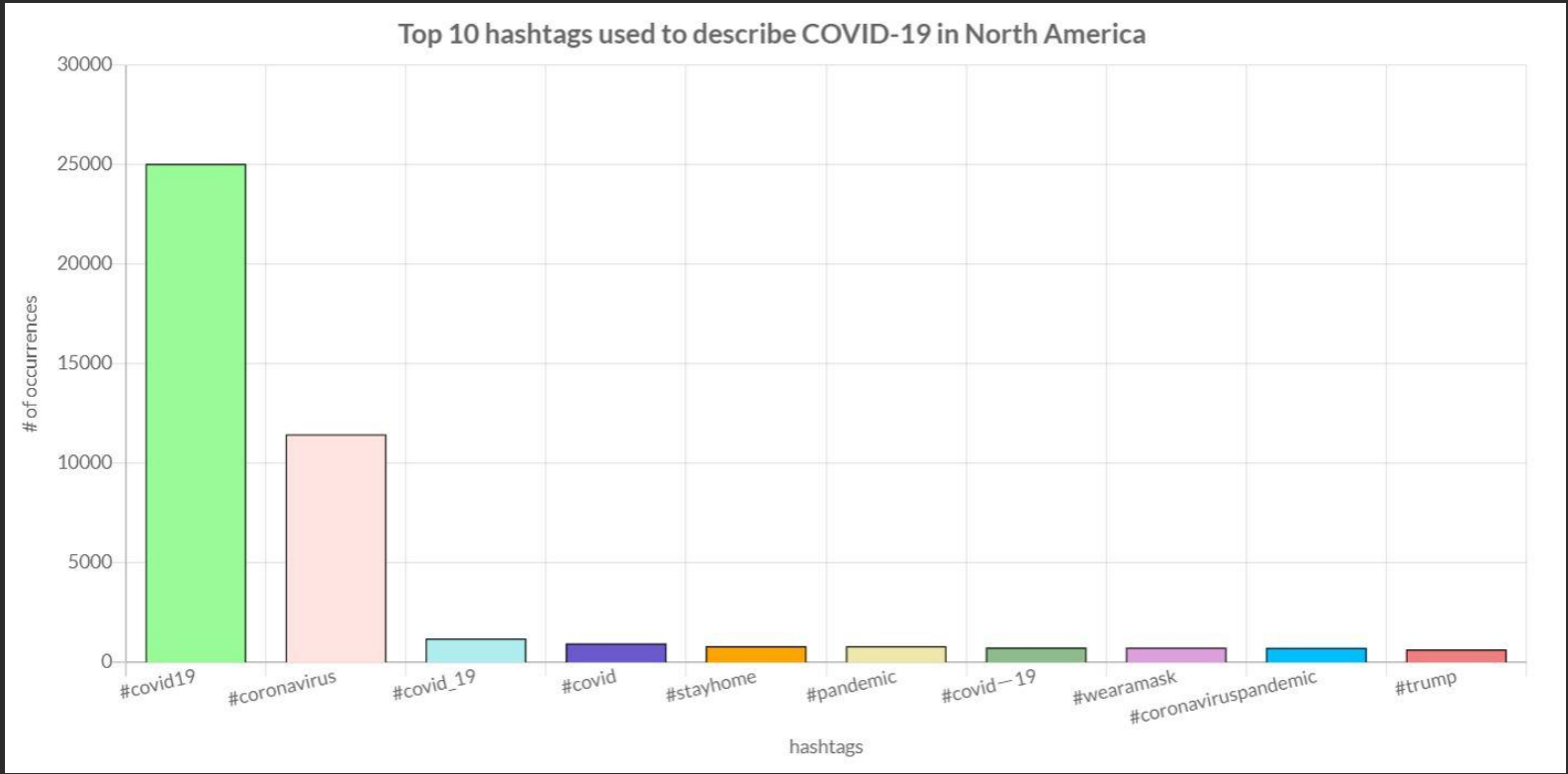
What are the hashtags used to describe COVID-19 by Region?



Question 7

Part A

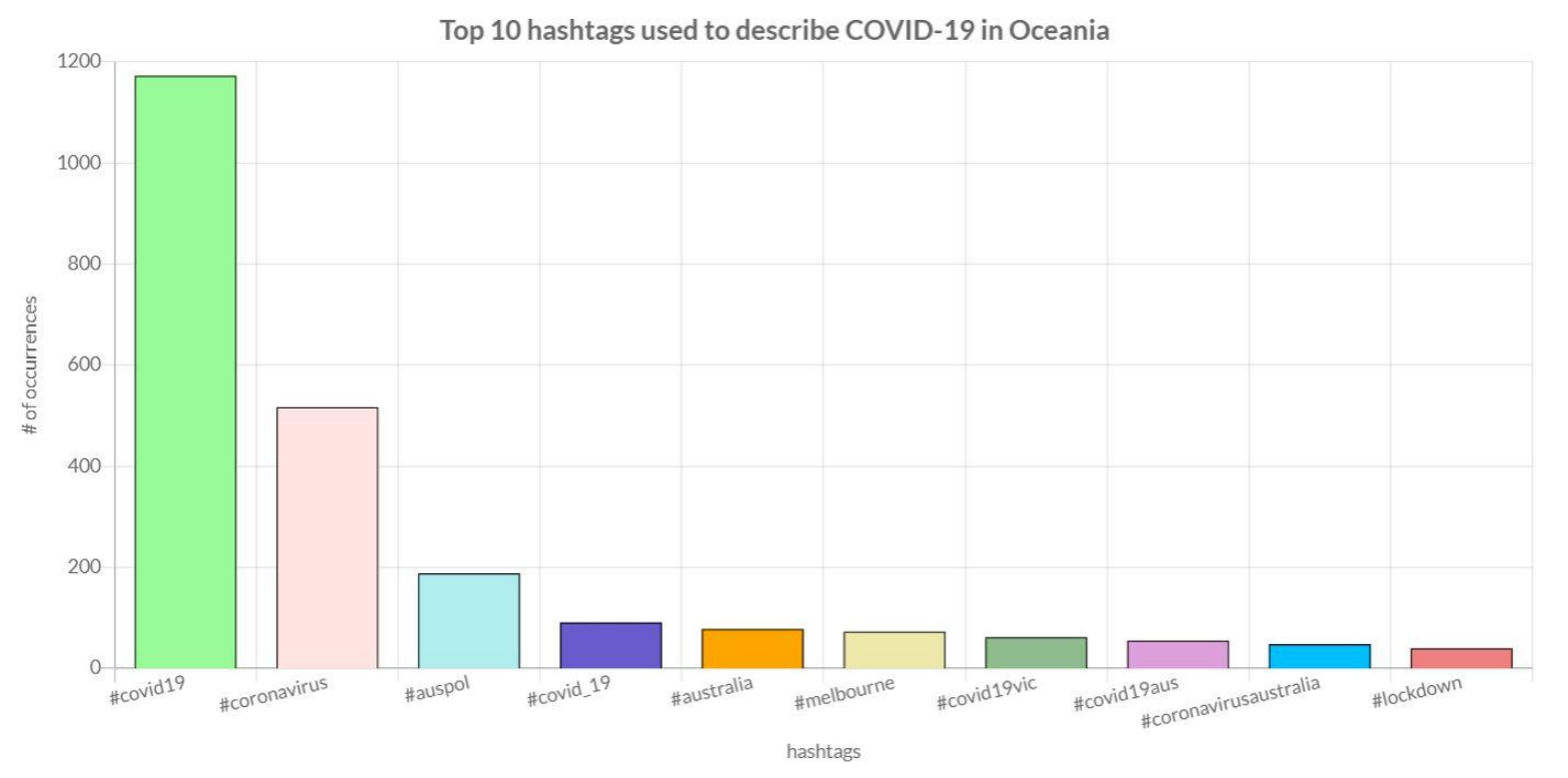
What are the hashtags used to describe COVID-19 by Region?



Question 7

Part A

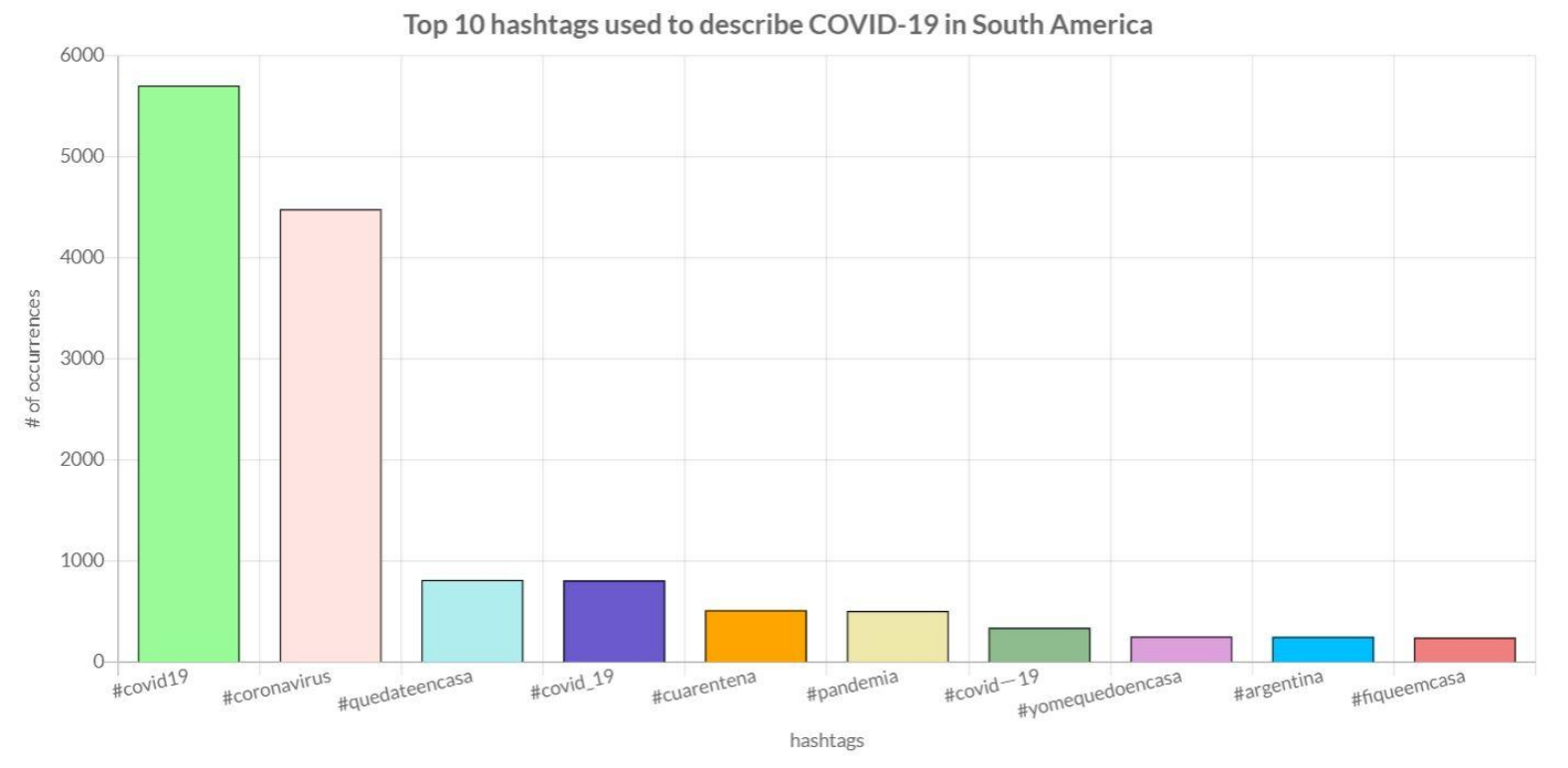
What are the hashtags used to describe COVID-19 by Region?



Question 7

Part A

What are the hashtags used to describe COVID-19 by Region?



Question 7

Part B

What are the top 10 commonly-used hashtags with COVID hashtags?

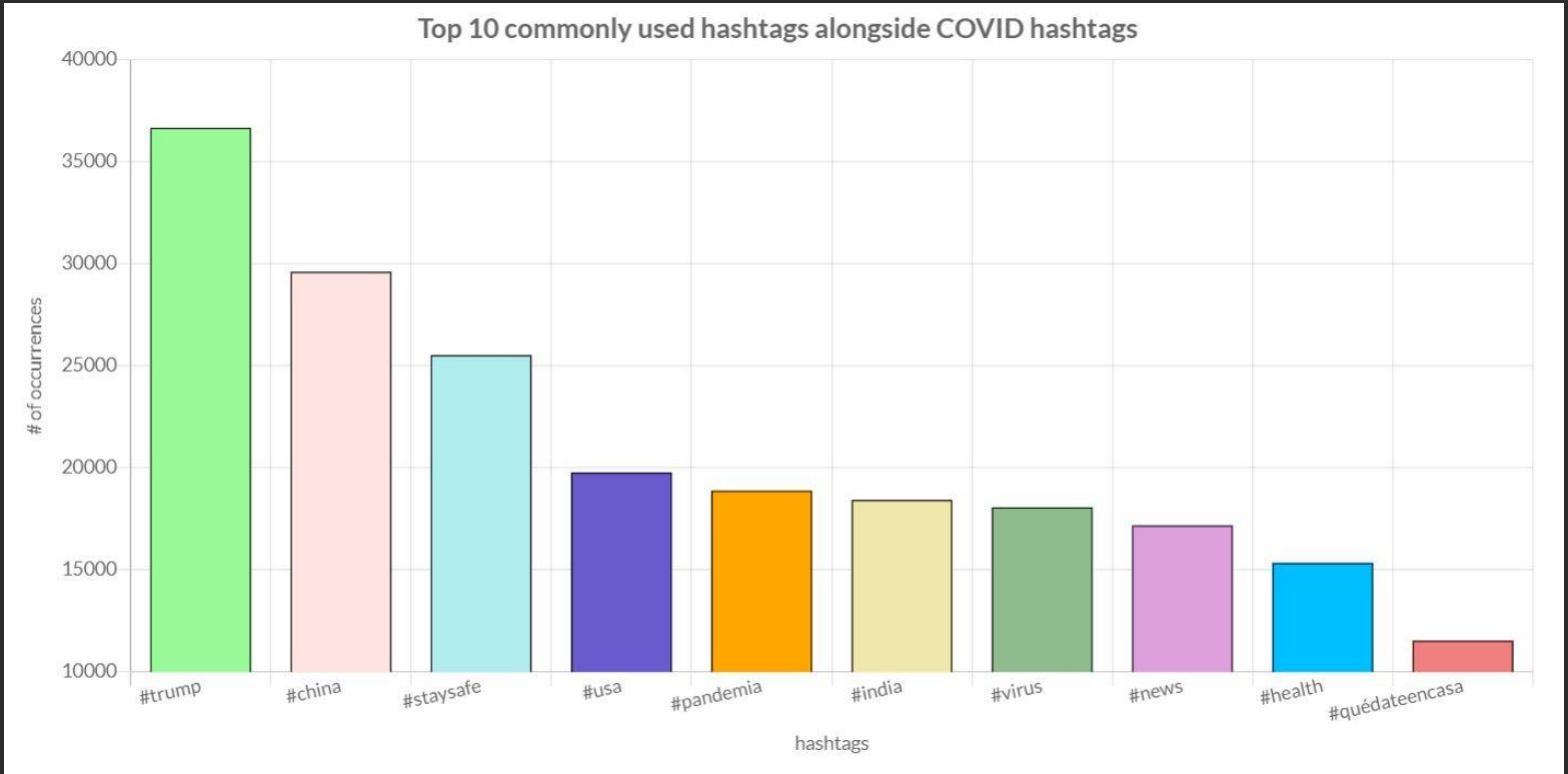
Overview of Process

- Receive the text data from tweet
- Filter out tweets with 0 or 1 hashtag and save tweets with 2 or more hashtags
- Create a list of hashtags related to COVID-19
  - Resource used: https://developer.twitter.com/en/docs/labs/covid19-stream/filtering-rules
- Filter to tweets that contain at least 1 hashtag from the COVID hashtag list
- Split tweets into individual words and perform word count
- Filter out all of the COVID hashtags from the COVID hashtag list and keep all other hashtags
- Perform group by and order by to obtain top 10 results

Question 7

Part B

What are the top 10 commonly-used hashtags with COVID hashtags?



Question 8

Blue Team: Liam Hood, D’Ante Jolly, Sean Tidd, Nahshon Williams

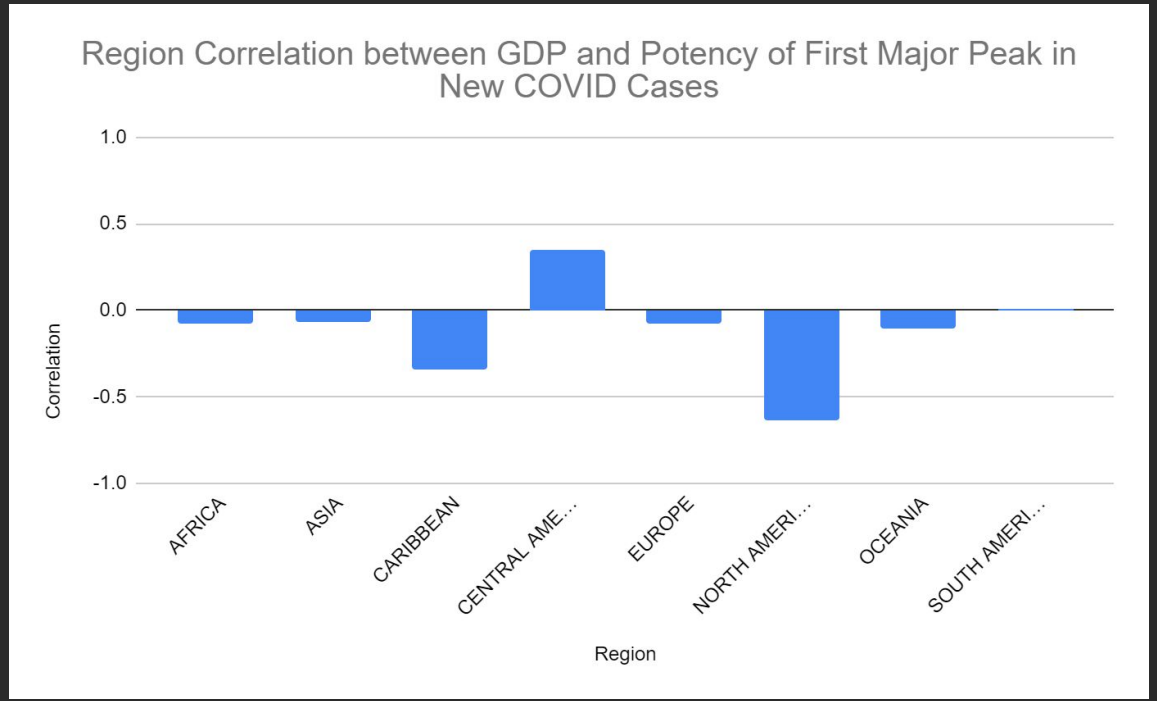
Objectives:

- Is there a relationship between a Region’s cumulative GDP and the intensity of the first spike in COVID cases?
- What is the average amount of time it took for each region to reach its first peak in infection rate per capita?

Question 8

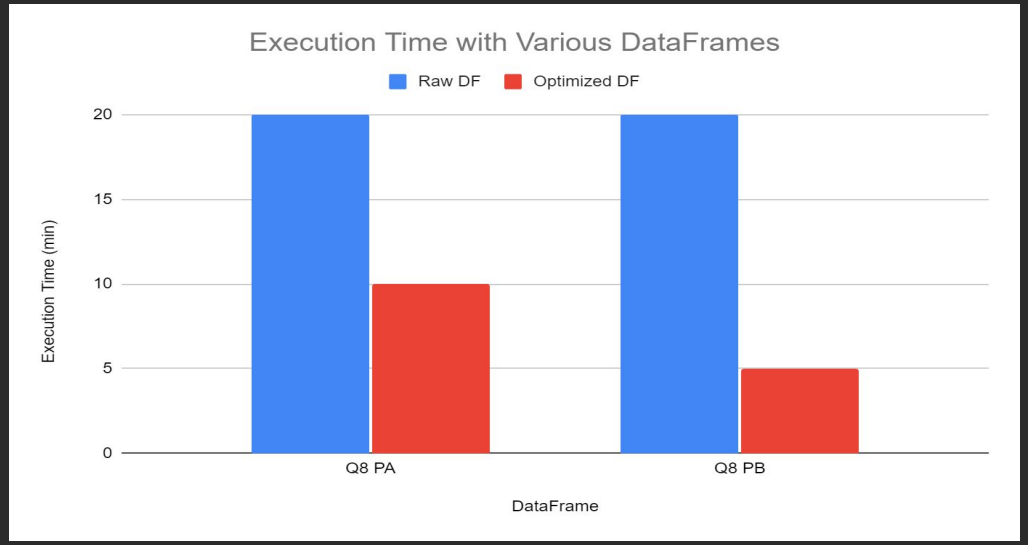
Part A

Is there a relationship between a Region's cumulative GDP and the first spike in COVID cases?



Question 8 - Optimization

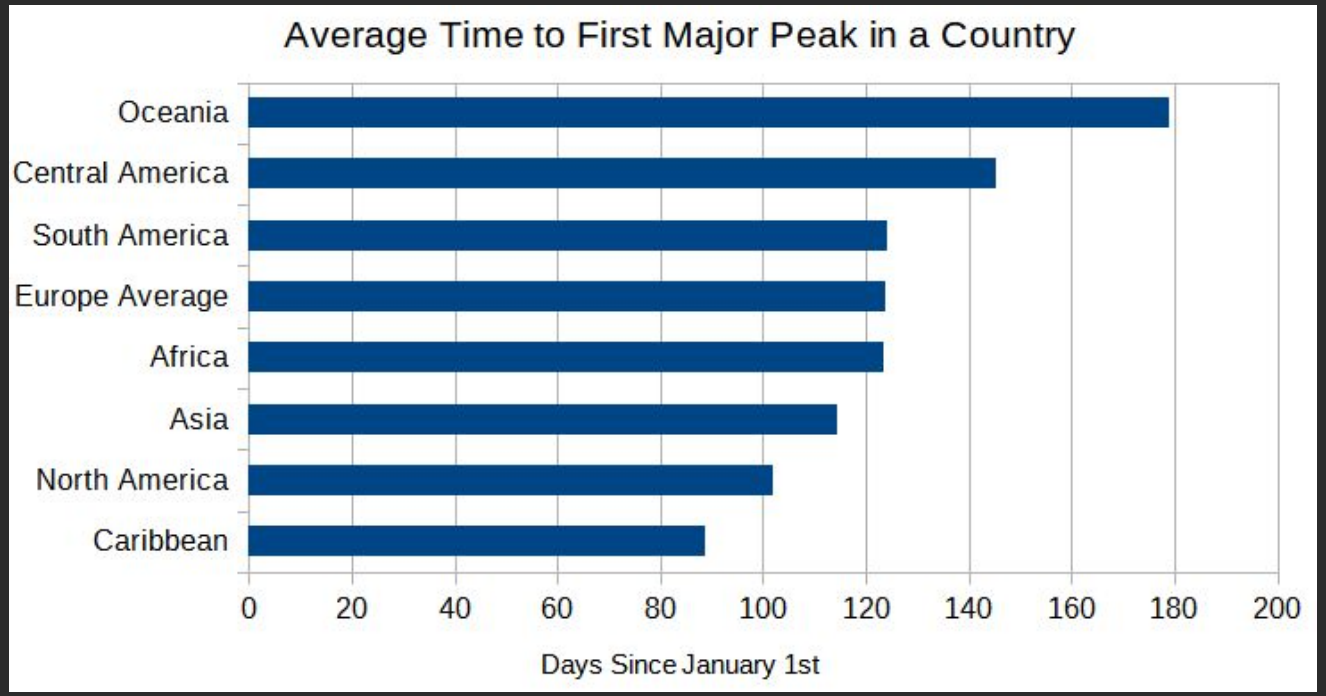
- Partitioned input DataFrame by regions (Africa, Asia, Caribbean, Central America, Europe, North America, Oceania, South America)
- Bucketed by 40 buckets for countries within each region partition
- Saved DataFrame as a Hive table to perform Spark SQL



Question 8

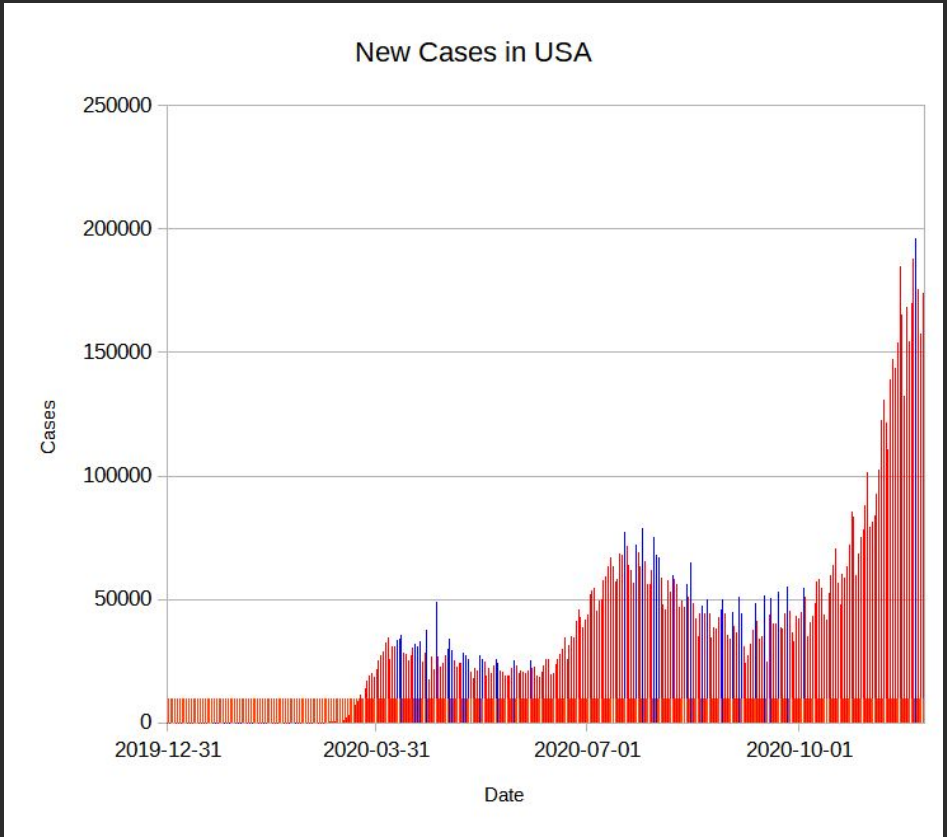
Part B

What is the average amount of time it took for each region to reach its first peak in infection?



Question 8 - First Peak Explanation

- Countries have clearer initial case spikes
- Peaks are local maximas in new cases, followed by an average decrease of 10 percent over the following 7 days
- Major peak is a peak that occurs after the new case count reaches 5 percent of the maximum number



Further Questions?