**General Findings**

1. [Age](https://www.kaggle.com/dkjung/covid-19-eda-s-korea-forecasting-global#eda_age)
   * 20s are the most infected age group (27.80% to total)
   * Older than 70 are the most deceased age group (48.08% to total)
   * Countries need to separate the two groups while preventing spreads as a whole
2. [Region](https://www.kaggle.com/dkjung/covid-19-eda-s-korea-forecasting-global#eda_location)
   * The more populated, the more cases in general (excluding 2 outliers)
   * What's riskier are community-level gatherings like religious events
   * Countries better try to prevent them and detect the particular event if it occured
3. [Sex](https://www.kaggle.com/dkjung/covid-19-eda-s-korea-forecasting-global#eda_sex)
   * Any significant difference between sexes was not found
   * Females seem more prone to be infected (58.91% VS. 41.09%)
   * Males seem more prone to be deceased (51.92% VS. 48.08%)
4. [Test](https://www.kaggle.com/dkjung/covid-19-eda-s-korea-forecasting-global#eda_test)
   * The number of tests and negative/released cases have kept increasing linearly
   * Confirmed/deceased cases show more dynamic slope changes than above three while increasing
   * Daily confirmed cases have the second spike around 2020-05-11 (a probable sign of the 2nd wave)
5. [Weather](https://www.kaggle.com/dkjung/covid-19-eda-s-korea-forecasting-global#eda_weather)
   * Weather-related features show no clear correlation with infections
   * There is a weak negative correlation between temperature and infections (colder & more infections)
   * There is a weaker positive correlation between maximun wind speed and infections (more wind & more infections)
6. [Patient](https://www.kaggle.com/dkjung/covid-19-eda-s-korea-forecasting-global#eda_patient)
   * Half of the logged patients have <=4 contacts with (or possible infections to) other patients before confirmed
   * There are outliers with more than 1,000 contacts though
   * The average number of contacts by a patient per day has decreased as time went by
7. [Seoul](https://www.kaggle.com/dkjung/covid-19-eda-s-korea-forecasting-global#eda_seoul) (the capital of S.Korea)
   * Floating population shows a positive correlation with the number of infections (more floating & more infections)
   * The activities of older than 60 decreased after the outbreak of COVID-19
   * The absolute number of cases in each subregion is too small (<100) to draw a conclusion from
8. [Search](https://www.kaggle.com/dkjung/covid-19-eda-s-korea-forecasting-global#eda_search)
   * Searching on pneumonia was prior to that on COVID-19
   * Increase of COVID-19 searches has been ahead of the actual increase of confirmed cases
   * The searching itself and the number of daily new cases have decreased (with a small spike around 2020-05-11)
9. [Policy](https://www.kaggle.com/dkjung/covid-19-eda-s-korea-forecasting-global#eda_policy)
   * Level 3 (2nd highest) caution on infectious diseases was implemented proactively
   * Policies on immigration have kept updated for more countries while the borders are open
   * Educational measures are yet shifting 'complete closure' > 'online classes' > 'physical classes (some graders)'

***Imagine a client who has asked you to do some analysis of the South Korean COVID-19 dataset. Who is the client and what have they asked you to do? Describe one maxim, question, or ethical commitment from the problem formulation phase of a data science project and apply it to this project. What would you do differently as a data scientist given your chosen maxim, question, or ethical commitment? Be prepared to illustrate your approach with examples drawn from the data.***

Client is the South Korean Ministry of Health and Welfare

Utilizing the information we have available in the data

* + For the most part appears to be 2 quarters worth, Jan 2020 through June 2020

Client interested in understanding whether any lessons can be learned from the data in terms of implementing COVID strategies for the future

**Question: How will the results of the analysis be used?**

Client: governmental agency that would be able to enact policies

* + Reopenings
  + Restrictions – border closings
  + Vaccination mandates

The analysis will have far reaching impacts across the full South Korea population. Some preliminary findings that could drive various policy decisions:

* + 20s are the most infected age group (27.80% to total)
  + Older than 70 are the most deceased age group (48.08% to total)
  + On analyzing the infection case, it can be observed that more number of patients are affected by coming in contact with the other patients followed by overseas inflow

Most infected group: do you shut down colleges and universities, restaurants, etc. to stop spread. How do you prevent exposure to an older population?

Similarly, do you decide to shut down borders to address one of the bigger infection causes which is overseas inflow

501 and 503 were both great in opening my eyes and understanding the impact data can disproportionately have on populations

* + Not everyone is affected equally – Regressive Power: those who need to go to in person jobs (restaurants, etc.) shutting down vs remote work
  + Accountability: analysis is your responsibility to ensure that it’s rolled out appropriately and does the least amount of harm to vulnerable populations. Everyone has an ethical responsibility

Important to consider that these two quarters may not be relevant to 2021 and moving forward (vaccine rollout).

**Do Differently:** Important to be the bridge between the analytics and the policy. No one knows the data better. Communicate directly with the client. Communicating results: have clear action steps

***Discuss the role of data visualization in data exploration and analysis. Be prepared to describe one basic and one advanced visual exploration technique that could be applied to the South Korean COVID-19 dataset. Explain how each visualization is encoded, how it is to be interpreted, and evaluate each in terms of its expressiveness and effectiveness if it was to be shared with a general audience. You should plan to experiment with different visualizations as part of your preparation for this question.***

Own day to day work and this data exploration – super valuable in data quality checks

* + Are there time gaps – what is the time range
  + Are there poorly populated variables – Null/unknown values appearing
  + Outlier evaluation – are there big spikes: useful for this type of dataset: time/weather related. Does it require cleaning or should these outliers be carried through?

Incredible benefits for visualization when discussing findings with stakeholders and clients

* + oftentimes complex information and data – can see/display clear relationships
  + provide an accessible way to see and understand trends (increase/decrease), outliers, and patterns in data.
  + Low barrier to entry – much easier to see a polished product of a heat map or histogram as opposed to a table that requires a viewer to digest the trend themselves
  + Can be manipulated but the goal would be to highlight the most important portions of the data – drawn to colors and patterns

**Basic**:

* + Time on X
    1. For the most part data is the first 2 quarters of 2020 (for the age, gender trending though really just Q2)
  + Confirmed cases, total count of people, deceased count, some value encoded for the Y axis
  + Encode the color of the line by gender, age groups, province
  + Encode marks by pointing to a spot on the line – can specify when policy changes ramped up. Level 1 – 3. Immigration policies

Expressive by showing trends and difference perhaps by gender, age, province and when changes were made to address increases. Effective graph in that it wouldn’t be cluttered, one there’s one clear visualization that establishes a relationship between two variables and how they change over time.

**Advanced**: Really enjoyed using geopandas in my own work in 521

* + Use longitude and latitude to get point geometry that could go into a Geo Data Frame from the case file that tracks the infection case
  + Can plot the points on South Korea itself to identify any clusters (perhaps from a day care, church event, etc.).
  + Marks can be used to signify why a particular cluster may have developed.

This would not be as expressive – no time series data, but effective in flagging clusters within the country and perhaps displaying why they occurred.

***We are all familiar with the maxim “correlation does not imply causation”. But what does it mean when we assert that two or more variables are statistically related? Describe the concept as well explain how to calculate correlation coefficients. Utilize the South Korean COVID-19 dataset to illustrate your points. Do associations exist (either positive or negative) between variables derived from the data? Is there a causal story to be told about the transmission, incidence, or persistence of the coronavirus in South Korea?***

Null hypothesis – no statistical significance between two variables

* + When we assert that two or more variables are statistically related it’s asserting that the relationship that is being exhibited in an analysis or finding did not happen simply by chance
  + Assessed by the p-value
    1. P-value tells you the likelihood of such a result happening if the null hypothesis was true
    2. Typically .05, so a 5% chance that you could have arrived at this finding with the assumption that the null hypothesis is correct
    3. Arbitrary – could be .01, .001, etc. to be as confident as possible. Sample size plays a factor
    4. Arbitrary also dangerous.
       1. P-hacking. Fishing for a relationship that may not be there
  + Statistically related goes beyond just observational. Testing the actual relationship as opposed to just noting an association and assuming that they’re linked

Correlation coefficients:

* + Positive or negative r squared
  + Closer to 1 or -1 is better. 0.2 is weak positive

**Relationships**

* + 20s are the most infected age group (27.80% to total)
  + Colder weather – more cases

This is not saying that 20 year olds have weak immune systems – more active at schools, going out, etc. so more likely to come in contact.

Join confirmed\_date and province in patient\_info to date and province in weather and you can assess average temperature and cases. Same with weather – COVID doesn’t travel better in hot or cold weather – cold weather brings people indoors and you’re more susceptible to exposure.

**Causal Story**: relationships exist but there driven by behavior moreso than an inherent relationship between COVID and the weather or someone in their 20s.

SPLOMs great way to find these relationships – easy through seaborn

***The split-apply-combine data analysis pattern, identified by Hadley Wickham in a*** [***2011 paper published in the Journal of Statistical Software***](https://www.jstatsoft.org/article/view/v040i01)***, describes a strategy for analyzing data. Describe the pattern and how it has been implemented in the Pandas library. Discuss how the split-apply-combine pattern can be utilized to glean insights from the South Korean COVID-19 dataset. Provide at least two examples of its use from your exploratory analysis of the data.***

Split Apply Combine

* + break up a big problem into small manageable pieces (Split),
  + operate on each piece independently (Apply)
  + put all the pieces back together (Combine).

Group By in SQL environments or python is the best way to accomplish this and is ubiquitous when using data exploration and analytic tools in pandas

Diagram

Description automatically generated

Boolean masking: limiting to a certain sub-section of a dataframe based on some column’s values

Filter: subsection of a dataframe’s rows and or columns

Resample: super helpful for time series data: if have 24 separate hours but want the day can sum up, or take a daily average

**Group By is the major focus**

* + Group by certain attributes and aggregate by sum, mean, median as some examples

20s are the most infected age group (27.80% to total)

Older than 70 are the most deceased age group (48.08% to total)

* + Can roll up the TimeAge file just by the age group category in a group by
  + Can resample the data too and assess over time

On analyzing the infection case, it can be observed that more number of patients are affected by coming in contact with the other patients followed by overseas inflow (PatientInfo File)

Power of MapReduce – mrjob. Ability to process large files and glean meaningful insights from it

***It's often said that data cleaning and manipulation takes up 80% of a data scientist's time, although that claim has recently been*** [***disputed***](https://blog.ldodds.com/2020/01/31/do-data-scientists-spend-80-of-their-time-cleaning-data-turns-out-no/)***.  No matter what the actual number is, describe the tools and/or techniques you utilize to make the data cleaning and manipulation effort more manageable, efficient, and scalable.  Use explicit examples from your analysis of the South Korean COVID-19 dataset.***

First step if possible – just open the file if Excel or CSV: easy way to quickly see some common pitfalls

* + Time series gaps
  + A lot of NaN: Detect outliers (through this process or just through a min/max/average check)
  + Non-Korean data from other countries in Patient Information

If that’s not possible due to the size of the file, a lot of techniques to understand what you’re using

* + Shape
  + Dtypes – character versus number in masking. Object in patient info is one annoying example

data.isnull().sum()

Can quickly see how many NaN records you have. So in the patient info file/dataframe there was 5,165 records and 1,122 cases where sex was blank and 1,300 where age was blank.

Not super ideal – important fields

* Dropna

df.dropna(subset=['age'])

* Fillna

A default value – clients have said if gender not available default to female. Large chunk blank so this is something to be wary of

* Merge (prefer in my own work having wide files easily aggregated for Tableau reporting etc).
  + Ease of use: join by date and have detail for a time period for age, gender, province, search and weather trends. Resample potentially required (day and hour, day and hour and gender, etc.

Spark really good for exploratory data analysis: fast and SQL based which is ideal for me and what I’m most familiar with

A picture containing table

Description automatically generated