GROUP ASSIGNMENT

TECHNOLOGY PARK MALAYSIA

AICT009-4-2-IDA

INTRODUCTION TO DATA ANALYTICS

UCDF2005(1)ICT(DI)

LECTURER: **Ms. Hema Latha Krishna Nair**

WEIGHTAGE: **80%**

TITLE: **Data Analysis on Global Covid-19 Impact**

GROUP: **Group 8**

PREPARED BY:

NAME	TP Number
LAI KAI YONG	TP059040

Workload matrix

Name	**Responsibilities**
Lai Kai Yong	• Scope: OLAP /BI • Additional Trend Line Analysis • Additional Clustering • Additional Forecasting
Kok Hon Kit	• Scope: OLAP /BI • Regression
Cheong Sheng Kui	• Scope: OLAP /BI • Clustering
Lim Wye Yee	• Scope: OLAP /BI • Classifier

Table of Contents

1.0 Introduction

Covid-19 is a serious virus pandemic that significantly impact many sectors not to mention economy. The huge changes in countries economics are big concerns to all nation especially the increase of inflation rate due to country lockdown which directly affect business closure. The inexperience of handling the pandemic impacts badly to the countries economy and had caused economic crisis. Until now, the Covid-19 virus outbreak is not being resolved fully as well as financial inflation in all countries. Recently, after more than a year of adaptation, many nations have moved to a recovery state where business are being operated as usual again. To proof the fact where countries economy is being affected throughout the period of the pandemic, a data analysis method should be implemented to get a clearer view and provide better suggestions for countries that had not proceed to the recovery state. In terms of existing analytics solution, many organizations such as International Monetary Fund (IMF), International Labour Organization (ILO) and Organization for Economic Co-operation and Development (OECD) had approached the pandemics with analytics method to assist in the planning of countries' economic recovery. However, each of the investigation and research are being implemented solely where economic indexes and labor force inspections are not compile together. Therefore, a data exploration on countries' economic indicators such as economic indexes, GDP changes, unemployment and employment should be conducted.

2.0 Business Aim

2.1 Goal

To explore the pattern of countries economy indicators changes due to the Covid-19 pandemic and provide insights towards countries economic growth.

2.2 Objectives

Scope 1: OLAP + BI on Countries Economic and Labor Force

Deliverables: Countries Economic Dashboard in Report

Definition	• Online Analytical Processing (OLAP) is a technology that support BI applications providing multidimensional analysis for data exploration (OLAP.com, 2021). • Business Intelligence (BI) is digitized process that analyzes data which assist in driving a better business decision (Stedman & Burns, 2020).
Aim	To identify countries economic problem and improve countries decision making in recovering their economies based on the insights gain from the dashboard.
Description	A dashboard that provides drilling up and drilling down function on location and time to view the data in more perspectives. With BI, the countries can plan and implement a better data-driven solution The capability to understand the data through different dimension can gain greater insights and maximize recovery efficiency.

Scope 2: Trend Line analysis on cases and economic index (Regression)

Deliverables: Trend Line Model on daily covid cases and daily economic indicator index

Definition	A trend line model describes a pattern on the direction of the data changes which get the best fit of the data (Chen, 2021). The trend line model is dependent on regression which is a predictive analysis technique indicating the relationship between a target variable and independent variable/variables (Ray, 2015).
Aim	To determine the trend on the changes of cases, economic indexes in the countries' economic indicators.

Description	The growth and changes on countries economic indexes and covid cases is being presented in a trend line modelling method. The trend may predict the future of the economic indexes based on the current and past information on these variables' value.

Scope 3: Clustering on Countries Economic Strength

Deliverables: Clustering Model on GDP Changes, Employment, Unemployment, Commodity Price Grouped by countries strength group

Definition	Clustering is an unsupervised learning that segment the data into groups where each of the data point in the same grouping are similar (Priy, 2021).
Aim	To indicate the meaningful clusters on countries based on their economics' strength affected by the covid pandemic.
Description	The intrinsic grouping among the unlabeled countries economic indicators data is being revealed by the clustering algorithm. The data points or countries that have similar data values on the economic indicators will be categorized based on certain assumption from the clustering method.

Scope 4: Time Series Forecasting on employment, unemployment and gdp changes

Deliverables: Time Series Forecasting predicting future economic indicators changes

Definition	Time Series Forecasting is a scientific approach to data in making prediction on future value based on the historical time stamped data (Tableau, 2021).
Aim	To predict the future growth of the countries' economic indicators value for better planning for countries to recover and improve their economy sector impacted by the pandemic crisis.
Description	GDP, employment and unemployment values are being predicted with the time horizon of yearly data through forecasting. The prediction values help countries to come up with better distribution on economic and greater planning for economical focus field.

3.0 Methodology

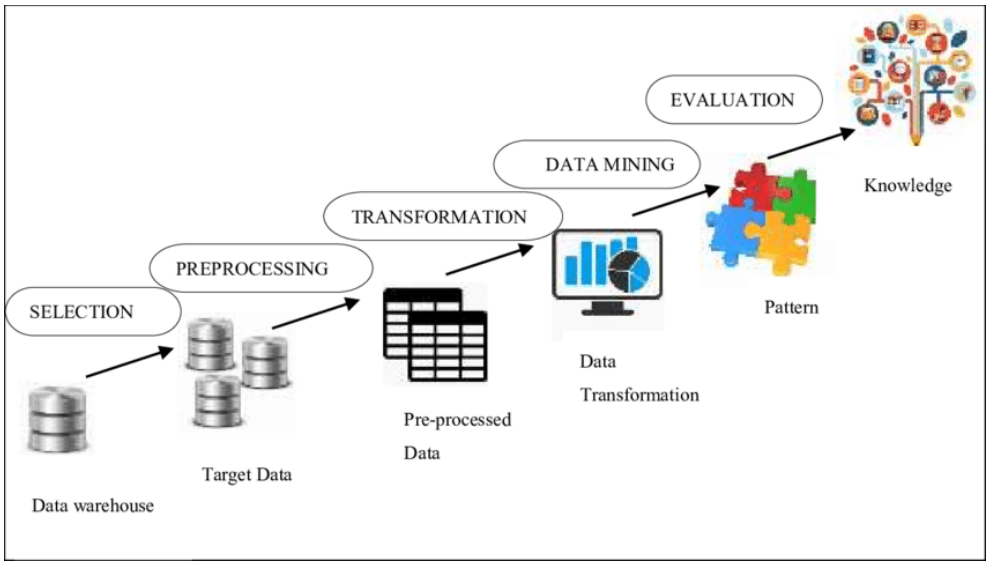


Figure 3.1: KDD Methodology
(Sabri, Man, Abu Bakar, & Rose, 2019)

The methodology chosen in diving into the country's economy data is Knowledge Discovery in Database (KDD). KDD is a popular data approaches process that is broadly used to explore and interpret knowledge from data by emphasizing pattern recognition, data visualization and machine learning technique (Hamilton, 2018). This methodology is selected as the identification on country's economy sector impact is required to include multiple datasets. It suits the need of choosing related columns or data variables and filter required timestamp for analysis. The relevant data selected is known as target data which is then being integrated in a data warehouse. With a data warehouse, all datasets are linkable where response of the visualized data is spontaneous based on the selection on slicer and drilling on the report. KDD requires researcher to select suitable data mining algorithm which suits the Covid-19 data-driven research in choosing relevant algorithm based on scope of research and datasets suitability. For countries economic analysis, supervised learning algorithms applied are regression and time series forecasting, meanwhile unsupervised learning method implied is clustering. In the analysis of the countries' economy, patterns and anomalies are being identified through the visualization and mining of data from the deliverables of OLAP/BI dashboard and machine learning models. The evaluation is then being performed on the determined pattern with the elaboration on real-world evidence and suggest recommendations.

Relating to each of the phases, the data warehouse is consisted of all datasets related to countries economy whenever relates to Covid-19 period without any filtering on details and merging in data tables. In this phase, the data for research is messy and noisy where null values are apparently have not being imputed. In the data selection process, unused data variables are being filtered and only desired datasets are selected out for further steps. Proceeding to data preprocessing, null values in data, wrong data types and irrelevant data indications are being handled to obtain a cleaned data. The cleaned data is then being passed through data transformation where in this phase. In data transformation, data are being restructured whenever required and data value modification. With a prepared and standardized data, the data mining is being implemented in descriptive analysis and predictive analysis. After having the deliverables like modelling and dashboard ready, the evaluation including analysis and recommendation is conducted to provide the countries with a data-driven suggestion in curbing the economic aftershock impacted by Covid-19.

4.0 Data Schema

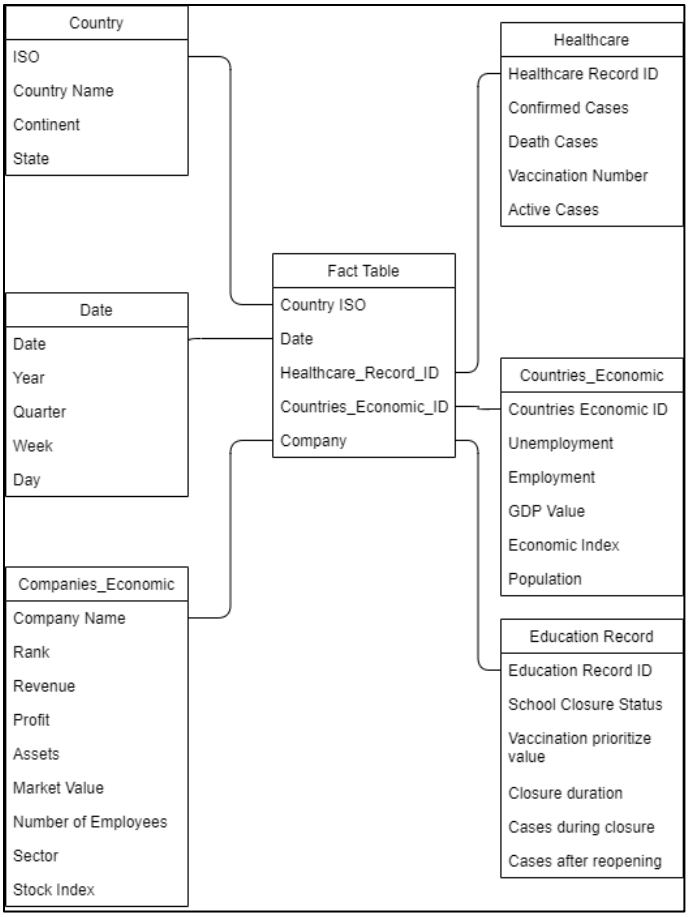


Figure 4.1: Domain Star Schema

The type of multidimensional data warehouse schema selected for the covid-19 data driven research is star schema. The dimension tables in the star schema are entities of the data warehouse which includes country, date, companies economic, healthcare, countries economic and education record. Each of the dimension tables hold a key column represent the data rows in the table which acts as unique record identifier. Key column in each dimension tables indicates the value of all followed descriptive columns. For example, each country ISO code in the country dimension table represents each row of data value for country name, continent, state and state fips. The schema has a fact table that acts as the connection point that created the multidimension of the data which contains key columns from all dimension tables. Focusing on countries economic analysis, the dimension tables used are country, date, healthcare (partial) and countries economic entities.

4.1 Data Selection

Data Selection is a crucial process in any data-focused research project. To explore the country's economy changes affected by Covid-19, multiple datasets had been selected to analyze the impact of Covid pandemics to countries' economy. The datasets used are covid healthcare data, countries GDP data, commodity price, economic indicators index, unemployment and employment data. Datasets are attained from the link listed below and categorized by main data columns in the datasets. All datasets are consisted of at least one column indicating respective country and are in time measurable basis.

Global Covid Data

- Our World in Data (OWID)
  https://github.com/owid/covid-19-data/tree/master/public/data
  Data Columns: Daily new cases, daily active cases, daily death cases + Economic Indexes

Countries' Economic Indicators Data

- International Monetary Fund
  https://www.imf.org/en/Publications/WEO/weo-database/2021/April/download-entiredatabase
  Data Columns: GDP, population, inflation, investment
- World Bank
  https://www.worldbank.org/en/research/commodity-markets#1
  Data Columns: Commodity price – crude oil, coal, energy, agriculture
  https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?end=2020&start=1994
  Data Columns: GDP Value

Labour force

- Organisation for Economic Co-operation and Development (OECD)
  https://data.oecd.org/unemp/unemployment-rate.htm
  Data Columns: Unemployment Value, Age Group, Education Level, Youth (Gender)

https://data.oecd.org/emp/employment-rate.htm
Data Columns: Employment Value, Age Group, Education Level, Activity, Working Hours

Data Description

Type	Field Name
#	Population
#	GDP Per Capita
Abc	ISO
Abc	Continent
Abc	Country
📅	Date
#	Total Cases
#	New Cases
#	Total Vaccination
#	People Vaccinated
#	People Fully Vaccinated
#	Stringency Index (SI)
#	Extreme Poverty
#	Human Development Index (HI

Figure 4.2: Partial of Data Variable Inspection

- Only ten selected data variables are discussed due to high volume of datasets

No.	Name	Type	Description
1	ISO	String	Country Code ISO 3166 Alpha-3
2	Country	String (Geographical)	Country Name
3	Continent	String	Continent Name
4	Date	Date	Date
5	New Cases	Integer (Whole)	New Covid-19 Cases daily
6	GDP Per Capita	Integer (Decimal)	Gross Domestic Product / Population
7	Extreme Poverty	Integer (Whole)	Extreme Poverty Reports Daily
8	GDP (USD)	Integer (Decimal)	Gross Domestic Product in Currency
9	Employment Value	Integer (Whole)	Observation value for employment
10	Population	Integer (Whole)	Countries' Population

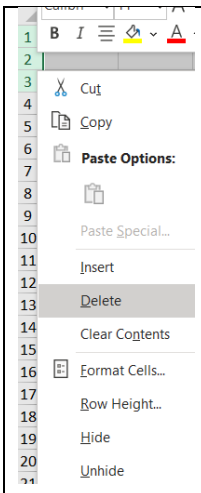
5.0 Data Pre-processing

Platform: *Excel*

Issue 1: Extra Upper Rows

It is identified that there are extra upper rows in the raw data which is not applicable for the identification of column headers in *Tableau Prep Builder*. Therefore, a process of deletion on unwanted upper rows is being conducted in order to place the data column headers on top of the spreadsheet. This enables *Tableau Prep Builder* to detect the column headers automatically without complicated action.

Implementation



1. **Select** the extra upper rows
2. **Right** click
3. Click **Delete** from menu option

Example 1

Before

	A	B	C	D	E	F	G
1	Data Source	World Development Indicators					
2	Last Updated Date	30/7/2021					
3							
4	Country Name	Country Code	Indicator Name	Indicator Code	1960	1961	1962
5	Aruba	ABW	GDP (current US$)	NY.GDP.MKTP.CD			
6	Africa Eastern and Southern	AFE	GDP (current US$)	NY.GDP.MKTP.CD	19291929320	19701861088	21470351753
7	Afghanistan	AFG	GDP (current US$)	NY.GDP.MKTP.CD	537777811.1	548888895.6	546666677.8
8	Africa Western and Central	AFW	GDP (current US$)	NY.GDP.MKTP.CD	10407321640	11131302981	11946843969
9	Angola	AGO	GDP (current US$)	NY.GDP.MKTP.CD			
10	Albania	ALB	GDP (current US$)	NY.GDP.MKTP.CD			
11	Andorra	AND	GDP (current US$)	NY.GDP.MKTP.CD			
12	Arab World	ARB	GDP (current US$)	NY.GDP.MKTP.CD			

After

	A	B	C	D	E
1	Country Name	Country Code	Indicator Name	Indicator Code	1960
2	Aruba	ABW	GDP (current US$)	NY.GDP.MKTP.CD	
3	Africa Eastern and Southern	AFE	GDP (current US$)	NY.GDP.MKTP.CD	19291929320
4	Afghanistan	AFG	GDP (current US$)	NY.GDP.MKTP.CD	537777811.1
5	Africa Western and Central	AFW	GDP (current US$)	NY.GDP.MKTP.CD	10407321640
6	Angola	AGO	GDP (current US$)	NY.GDP.MKTP.CD	
7	Albania	ALB	GDP (current US$)	NY.GDP.MKTP.CD	

Example 2

Before

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	**World Bank Commodity Price Data (The Pink Sheet)**												
2	**monthly prices in nominal US dollars, 1960 to present**												
3	**(monthly series are available only in nominal US dollars)**												
4	Updated on September 02, 2021												
5		Crude oil, average	Crude oil, Brent	Crude oil, Dubai	Crude oil, WTI	Coal, Australian	Coal, South African	Natural gas, US	Natural gas, Europe	Liquefied natural gas, Japan	Natural gas index	Cocoa	Coffee, Arabica
6		($/bbl)	($/bbl)	($/bbl)	($/bbl)	($/mt)	($/mt)	($/mmbtu)	($/mmbtu)	($/mmbtu)	(2010=100)	($/kg)	($/kg)
726	2019M11	60.40	62.74	61.41	57.06	66.99	73.62	2.63	5.15	10.04	63.27	2.52	3.11
727	2019M12	63.35	65.85	64.41	59.80	66.18	76.03	2.20	4.62	10.06	55.55	2.44	3.46
728	2020M01	61.63	63.60	63.76	57.52	69.66	82.09	2.02	3.63	9.89	48.52	2.60	3.13
729	2020M02	53.35	55.00	54.51	50.53	67.64	79.99	1.90	2.91	9.89	43.61	2.72	2.99
730	2020M03	32.20	32.98	33.75	29.88	66.74	67.89	1.78	2.72	10.21	41.42	2.34	3.27
731	2020M04	21.04	23.34	23.27	16.52	58.55	56.58	1.73	2.12	10.01	37.90	2.27	3.41

After

	A	B	C	D	E	F	G	H	I	J	K
1		Crude oil, average	Crude oil, Brent	Crude oil, Dubai	Crude oil, WTI	Coal, Australian	Coal, South African	Natural gas, US	Natural gas, Europe	Liquefied natural gas, Japan	Natural gas index	Co
2		($/bbl)	($/bbl)	($/bbl)	($/bbl)	($/mt)	($/mt)	($/mmbtu)	($/mmbtu)	($/mmbtu)	(2010=100)	($/
4	1960M01	1.63	..	1.63	..	..	..	0.14	0.40	..	3.62	
5	1960M02	1.63	..	1.63	..	..	..	0.14	0.40	..	3.62	
6	1960M03	1.63	..	1.63	..	..	..	0.14	0.40	..	3.62	
7	1960M04	1.63	..	1.63	..	..	..	0.14	0.40	..	3.62	
8	1960M05	1.63	..	1.63	..	..	..	0.14	0.40	..	3.62	
9	1960M06	1.63	..	1.63	..	..	..	0.14	0.40	..	3.62	
10	1960M07	1.63	..	1.63	..	..	..	0.14	0.40	..	3.62	

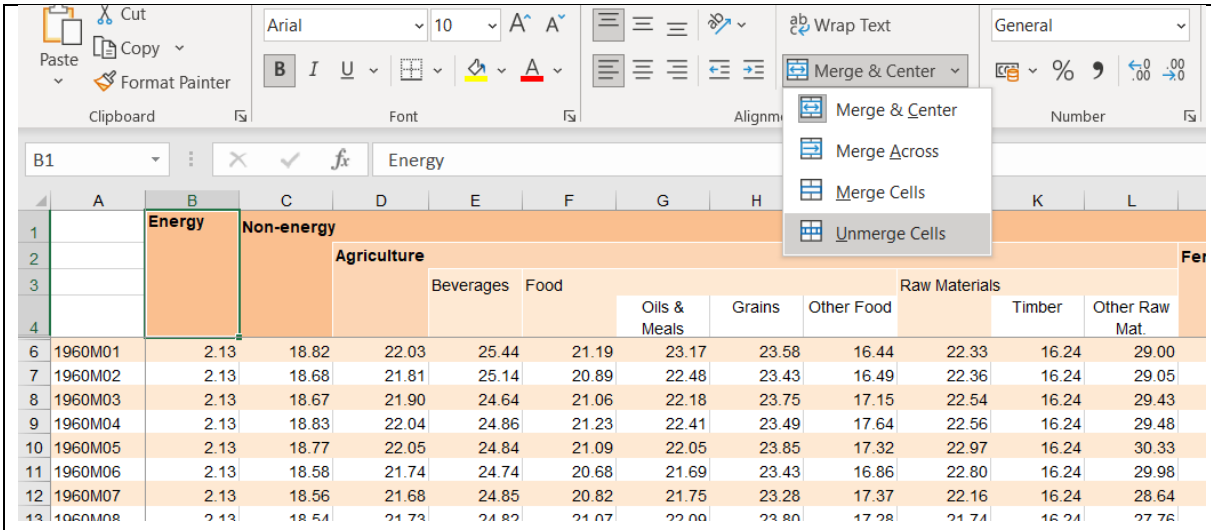
Note:

- Freeze Pane will be ignored in *Tableau Prep Builder* and data is served based on columns and row value.
- An original file of data is being remain and amendment is applied to a new file.

Issue 2: Merge Field

Merge fields are spotted in the raw data which is specifically applied on the column headers. The merge fields should be flattened for the simplification on the *Tableau Prep Builder* to identify the data headers. The action of flattening includes unmerging the cells and removing any unwanted rows to standardize the data header to exactly one row.

Implementation



1. **Select** the merged data field
2. Click **Unmerge Cells** in the Home tab (Alignment Section)
3. **Remove** extra rows after all unmerge action which flatten the headers to a row.

Example

Before

		Energy	Non-energy													
				Agriculture									Fertilizers	Metals & Minerals		Precious Metals
					Beverages	Food				Raw Materials					Base Metals (ex. iron ore)	
							Oils & Meals	Grains	Other Food		Timber	Other Raw Mat.				
6	1960M01	2.13	18.82	22.03	25.44	21.19	23.17	23.58	16.44	22.33	16.24	29.00	12.86	12.90	14.07	3.27
7	1960M02	2.13	18.68	21.81	25.14	20.89	22.48	23.43	16.49	22.36	16.24	29.05	12.86	12.93	14.11	3.27
8	1960M03	2.13	18.67	21.90	24.64	21.06	22.18	23.75	17.15	22.54	16.24	29.43	12.86	12.72	13.86	3.27
9	1960M04	2.13	18.83	22.04	24.86	21.23	22.41	23.49	17.64	22.56	16.24	29.48	12.86	12.93	14.12	3.27
10	1960M05	2.13	18.77	22.05	24.84	21.09	22.05	23.85	17.32	22.97	16.24	30.33	12.86	12.73	13.87	3.27
11	1960M06	2.13	18.58	21.74	24.74	20.68	21.69	23.43	16.86	22.80	16.24	29.98	12.86	12.75	13.90	3.27
12	1960M07	2.13	18.56	21.68	24.85	20.82	21.75	23.28	17.37	22.16	16.24	28.64	12.86	12.81	13.96	3.27
13	1960M08	2.13	18.54	21.73	24.82	21.07	22.09	23.80	17.28	21.74	16.24	27.76	12.86	12.65	13.77	3.27
14	1960M09	2.13	18.59	21.88	24.72	21.31	22.19	24.14	17.58	21.82	16.24	27.93	12.86	12.50	13.59	3.27

After

	Energy	Non-energy	Agriculture	Beverages	Food	Oils & Meals	Grains	Other Food	Raw Materia	Timber	Other Raw Mat.	Fertilizers	Metals & M	Base Metals (ex. iron ore)	Precious Metals
1960M01	2.13	18.82	22.03	25.44	21.19	23.17	23.58	16.44	22.33	16.24	29.00	12.86	12.90	14.07	3.27
1960M02	2.13	18.68	21.81	25.14	20.89	22.48	23.43	16.49	22.36	16.24	29.05	12.86	12.93	14.11	3.27
1960M03	2.13	18.67	21.90	24.64	21.06	22.18	23.75	17.15	22.54	16.24	29.43	12.86	12.72	13.86	3.27
1960M04	2.13	18.83	22.04	24.86	21.23	22.41	23.49	17.64	22.56	16.24	29.48	12.86	12.93	14.12	3.27
1960M05	2.13	18.77	22.05	24.84	21.09	22.05	23.85	17.32	22.97	16.24	30.33	12.86	12.73	13.87	3.27
1960M06	2.13	18.58	21.74	24.74	20.68	21.69	23.43	16.86	22.80	16.24	29.98	12.86	12.75	13.90	3.27
1960M07	2.13	18.56	21.68	24.85	20.82	21.75	23.28	17.37	22.16	16.24	28.64	12.86	12.81	13.96	3.27
1960M08	2.13	18.54	21.73	24.82	21.07	22.09	23.80	17.28	21.74	16.24	27.76	12.86	12.65	13.77	3.27
1960M09	2.13	18.59	21.88	24.72	21.31	22.19	24.14	17.58	21.82	16.24	27.93	12.86	12.50	13.59	3.27
1960M10	2.13	18.11	21.22	24.79	20.32	21.46	23.24	16.18	21.59	16.24	27.44	12.86	12.32	13.37	3.27

Note:

- The alignment of data value in the cells are ignored since it will not affect the data import process.

5.1 Data Import

Platform: *Tableau Prep Builder*

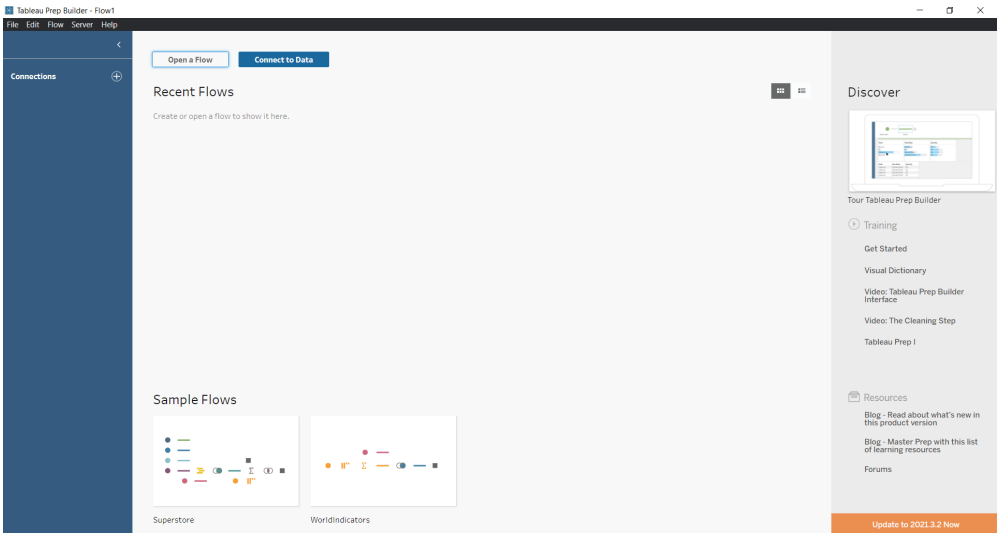
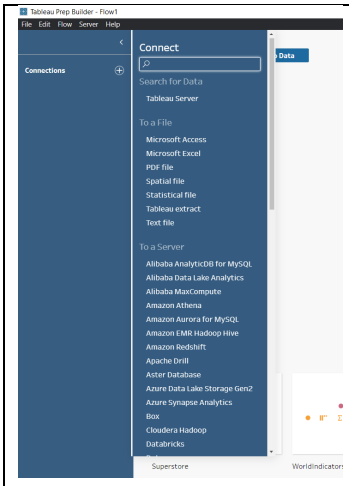


Figure 5.1: Tableau Prep Builder

To create a data warehouse, the first step is to import all datasets to a data preparation platform. In this case, *Tableau Prep Builder* is being utilized. Considering the different type of file format which the data is being stored, the data import can be performed by connecting to either Microsoft Excel file (.xlsx, .xls) or Text file (.csv).

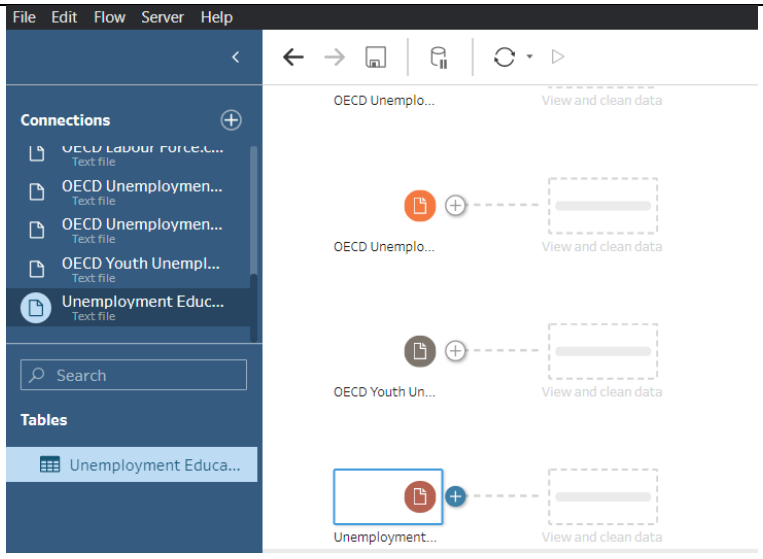
Implementation



1. Click **Connect to Data**
2. Select **File Type**
3. Open **Desired Datasets** in platform

Result

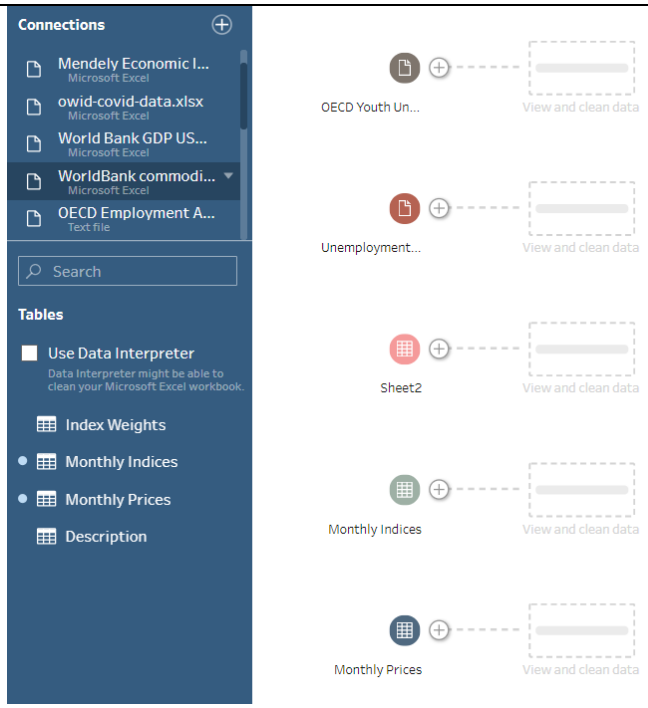
After



After importing the total of eleven datasets, the selection on required tables is being performed. For spreadsheets that only have one sheet, it is automatically being selected by the application. The selection on required tables is only applied to datasets that have more than one sheet / table in the file.

Example

After



5.2 Data Understanding

Platform: *Tableau Prep Builder*

Icon	Data type
Abc	Text (string) values
	Date values
	Date & Time values
#	Numerical values
T|F	Boolean values (relational only)
	Geographic values (used with maps)
	Cluster Group (used with Find Clusters in Data)

Figure 5.2： *Tableau* Supported Data Type
(Tableau, 2021)

Before cleaning the data, data understanding is very important. According to the Tableau official webpage, the data types supported are text, date. Datetime, integer, Boolean, geographical and cluster group. Whereas, in the process of data cleaning and data transformation, the data understanding has a direct impact to the accuracy and relevancy of the end result. Based on the data imported, there are one healthcare Covid-19 cases table, seven labor force related data tables and four economy data tables. By looking through the data, there are certain terminologies that are required to understand before diving into data.

1. Stringency Index (SI)

   Definition: SI indicates the strictness of standard operation policies (specifically on lockdown) which restrict citizen's behavior. The index value is calculated through public information events, ordinal containment and closure policy measures (Blavatnik School of Government, 2021).
2. Human Development Index (HDI)

Definition: HDI measures the key dimension of human development including life healthiness, knowledgeability and living standard. HDI can emphasizes on people's capabilities in the assessment of country development often applied in the planning of nationwide policies and express gross national income (UNDP, 2021).

3. Gross Domestic Product (GDP)

   Definition: GDP is a measurement on financial value of goods and services towards end users produced in a country in a specific timeframe. It is commonly used in defining the health of a nation or global in terms of monetary state (Callen, 2020)

6.0 Data Preparation

In data preparation, both data cleaning and data transformation is being implemented to obtain a well-structured data for further analysis.

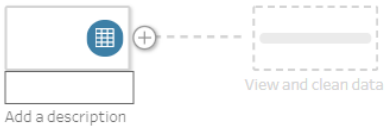
Platform: *Tableau Prep Builder*

Task 1: Naming Convention

Naming convention on data tables and data column is very important to better identify and recognize the data records. This avoids any misplacement or misuse of the data records especially during the data integration which requires joining of the data key column.

- Table Name

Implementation

	1. **Double-click** on the existing table name 2. Insert appropriate table name 3. Add **description** (source, summary)

While changing the name of data tables, a description is added to each of the data tables for better recognition on the source of datasets and summary on the table.

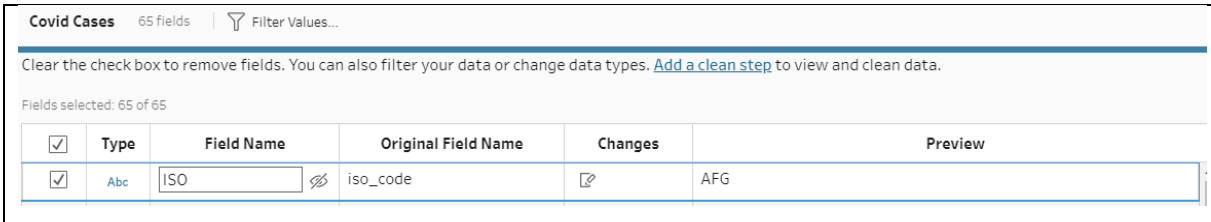
Result (Partial)

Before	After
	

- Column Name

Column names are being standardized and rename with appropriate text. For example, all country code is being standardized to ISO as the column name instead of having code, . The reason of having naming convention on columns name is to avoid any confusion and having a better name of data representation while visualizing the data on statistical graphics. Moreover, most of the column names are being capitalized so it looks better in terms of data presentation.

Implementation



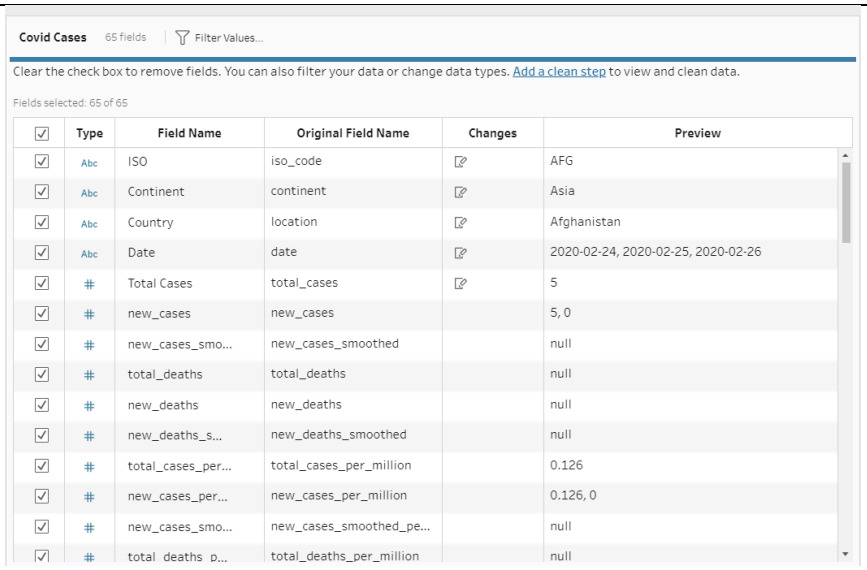
1. **Click** on Field Name
2. **Type** in desired column name

Note:

- Naming convention only applied to desired data variables for analysis.

Result (Partial)

After

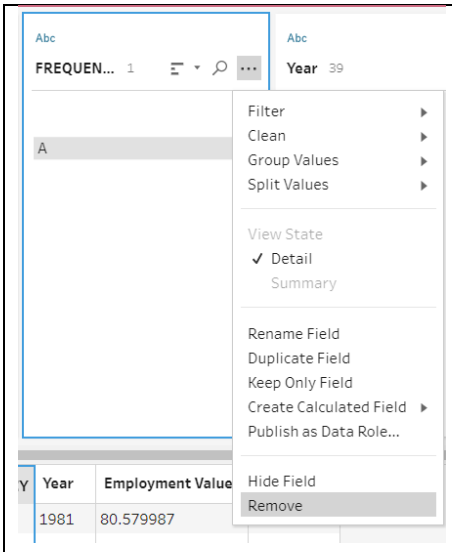


Type	Field Name	Original Field Name	Changes	Preview
Abc	ISO	iso_code		AFG
Abc	Continent	continent		Asia
Abc	Country	location		Afghanistan
Abc	Date	date		2020-02-24, 2020-02-25, 2020-02-26
#	Total Cases	total_cases		5
#	new_cases	new_cases		5, 0
#	new_cases_smo...	new_cases_smoothed		null
#	total_deaths	total_deaths		null
#	new_deaths	new_deaths		null
#	new_deaths_s...	new_deaths_smoothed		null
#	total_cases_per...	total_cases_per_million		0.126
#	new_cases_per...	new_cases_per_million		0.126, 0
#	new_cases_smo...	new_cases_smoothed_pe...		null
#	total_deaths_p...	total_deaths_per_million		null

Task 2: Exclude Unwanted Columns

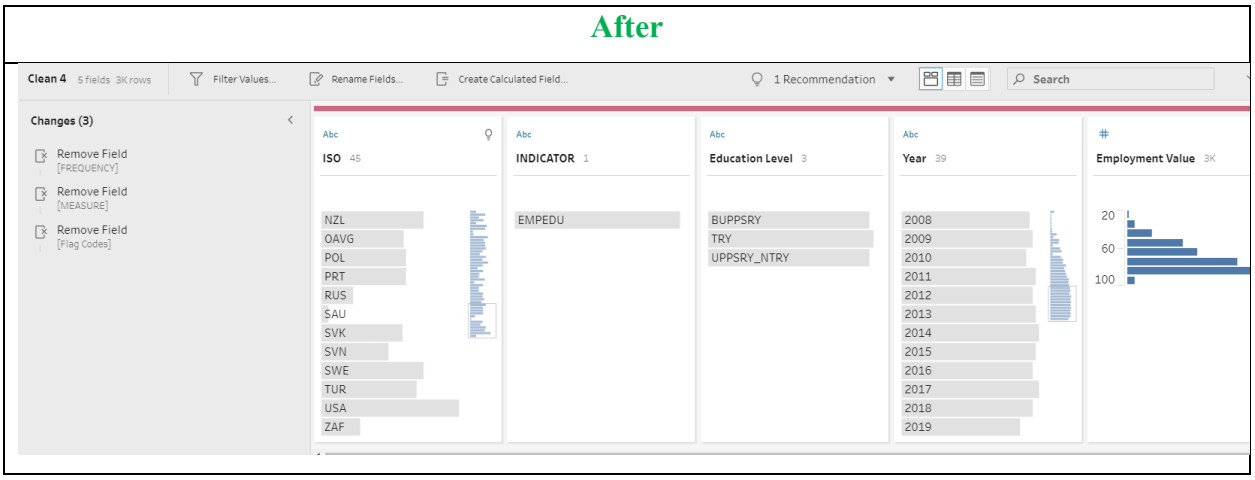
To ensure the performance and reduce any unwanted wastage on processing the data in both *Tableau Prep Builder* and *Tableau Desktop*, unwanted data columns are removed / excluded.

Implementation



1. Click the **triple dots** beside the column
2. Select **Remove**

Result (Partial)



Task 3: Null Value Handling

- **Scenario 1:** Exclude Null Rows (Special Case)

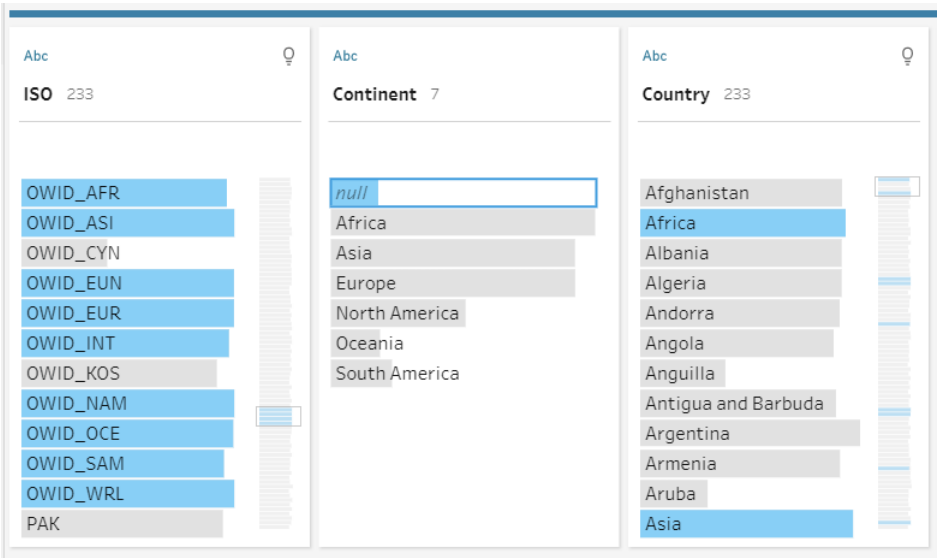


Figure 6.1: NULL Value in Covid Cases Data Continent

In the Covid-19 cases dataset, there are null values in the Continent column where they actually able to be aggregated just with the continent values that are present. The reason of having this statement is that looking into the ISO and Country columns, the values are actually the name

of continents and total of the world covid related information. Hence, the removal of the null values in Continent is a valid choice.

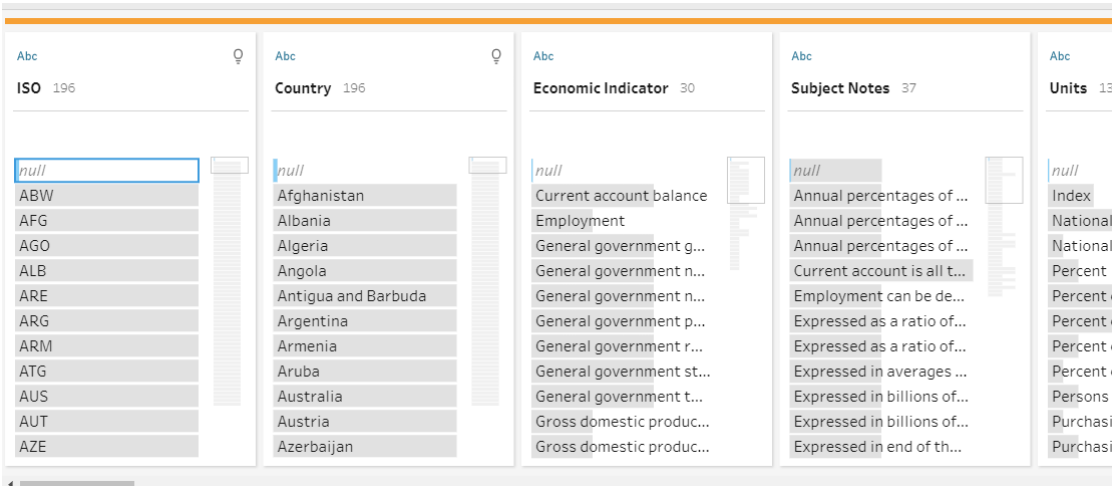
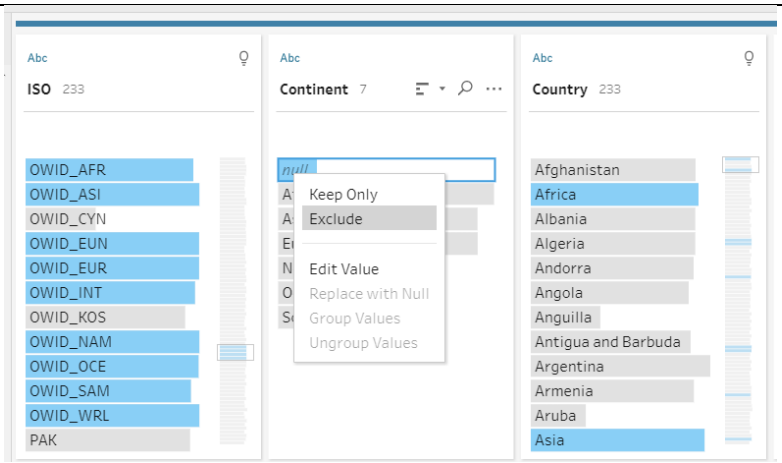


Figure 6.2: NULL – Empty Rows in Economic Overview Data

In the Economic Overview and Commodity Price data, there are null values that are empty rows. Regarding this scenario, removal of null values is acceptable as well. Regardless the usage of data removal is valid in this case, this data cleaning way is still not a good option since it will cause a data loss in the data research.

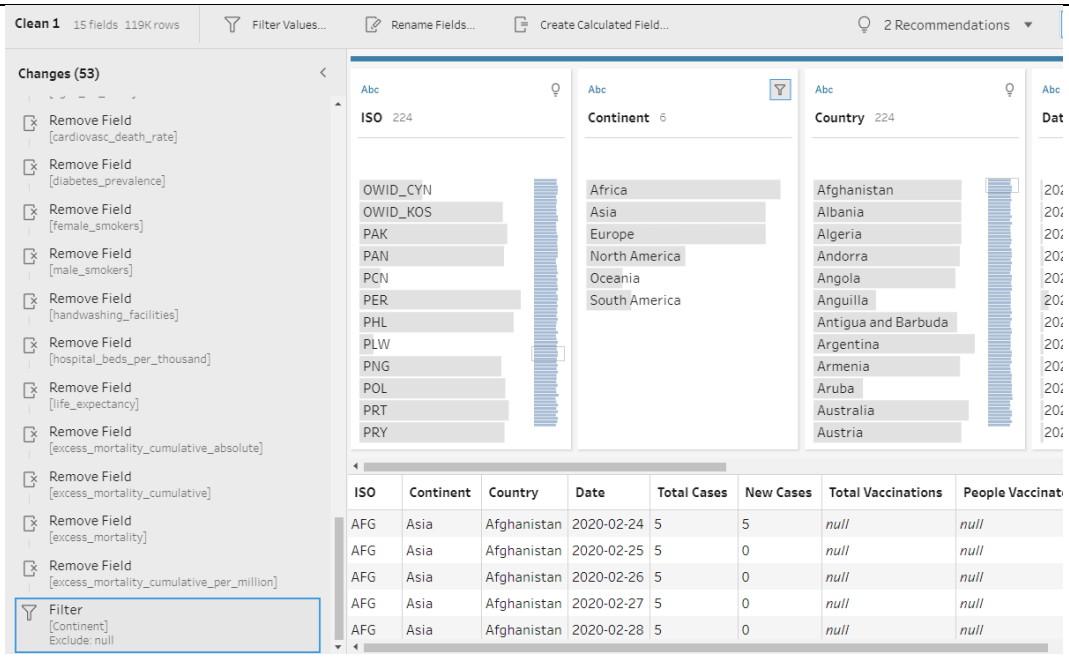
Implementation (Partial)



1. **Right Click** on selected null values
2. Select **Exclude**

Result (Partial)

After



- **Scenario 2:** Replace with 0

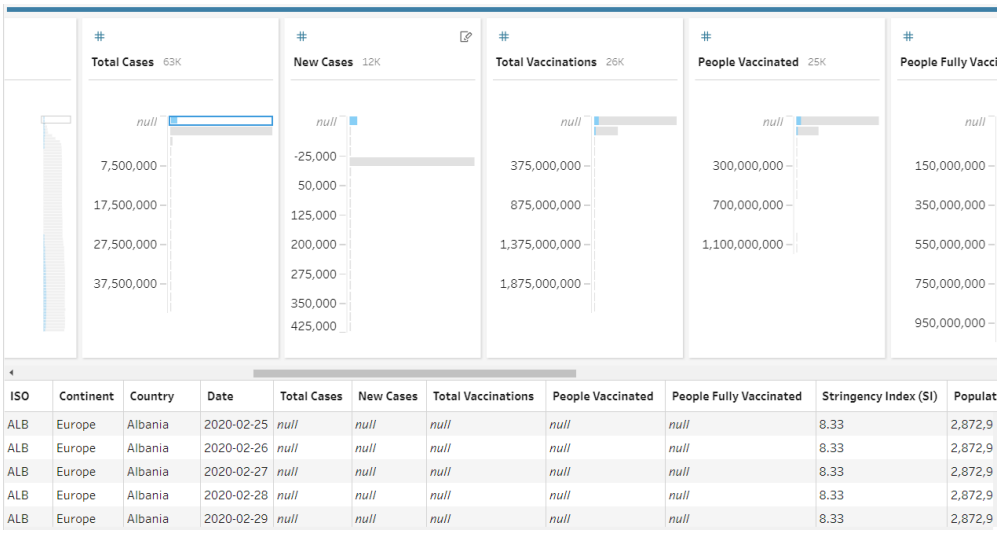
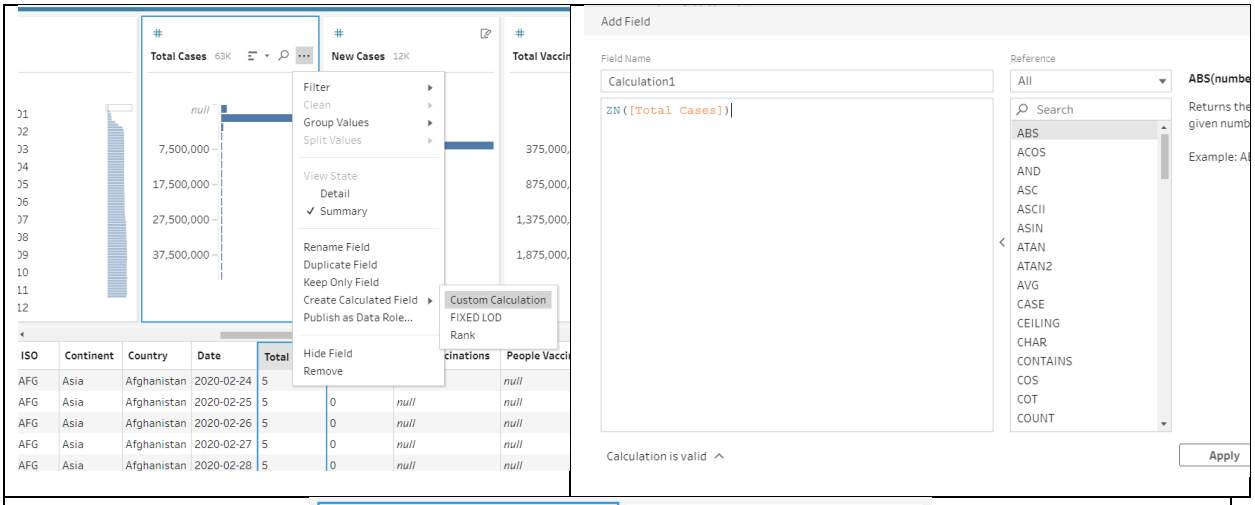
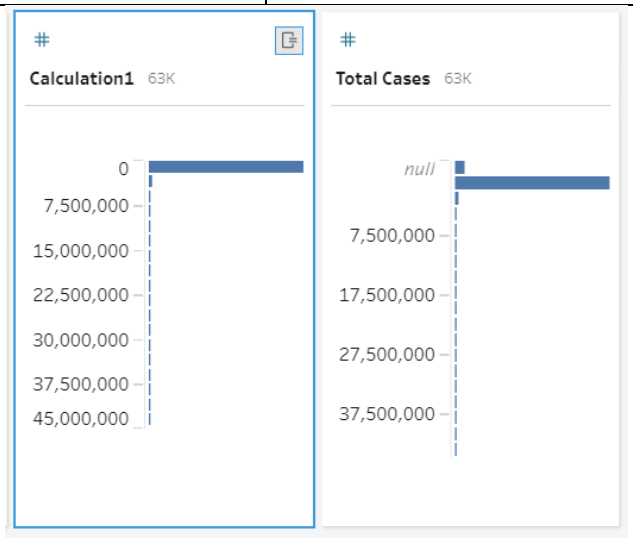


Figure 6.3: NULL Value in Covid Cases Data Total Cases

In the Covid Cases data, there are also null values in cases and vaccination related columns. However, based on the preview of the selected null value records, it is logically acceptable. This is because at the meantime, the resulted countries have not faced any Covid-19 infection yet and vaccines are not available in the meantime. Consequently, the null values should be replaced with zero which is a better representation of data rather than leaving a blank cell. From the perspective of stringency index, the SOPs and lockdown had not begun so zero is obviously better than null.

Implementation (Partial)

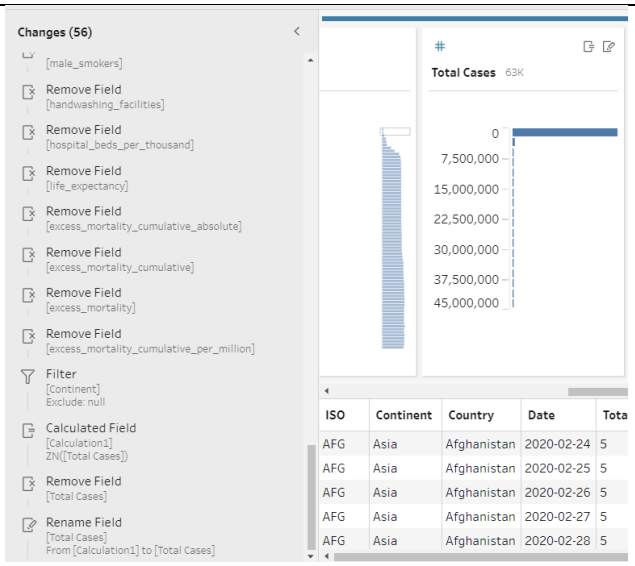




1. Click the **triple dots** on the column.
2. Choose **Create Calculated Field > Custom Calculation**
3. Insert `FN([COLUMN NAME])` and apply.
4. **Drop** the original column and **rename** the new calculation field.

Result (Partial)

After



ZN function is a popular null value handling method in Tableau by replacing zero for null values and return the expression when it is stored with value (Tableau, 2021).

- **Scenario 3:** Replace with Mean (Data Imputation)

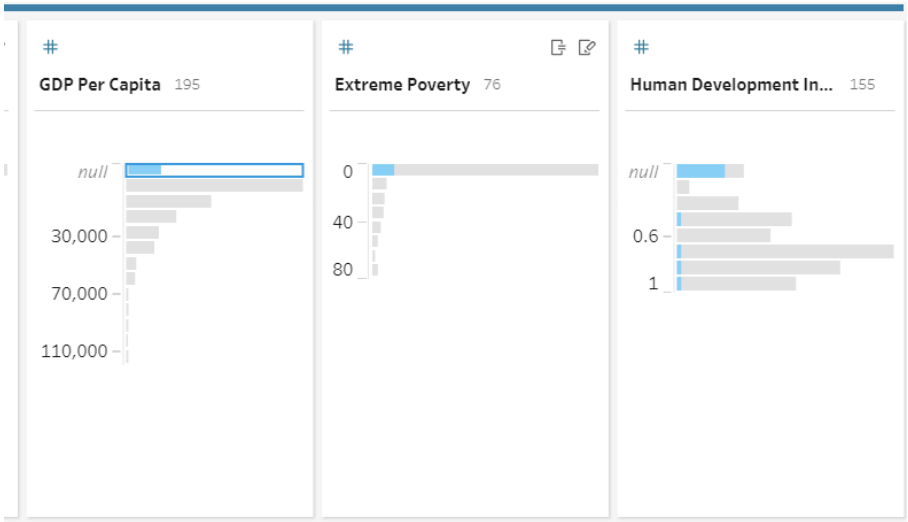
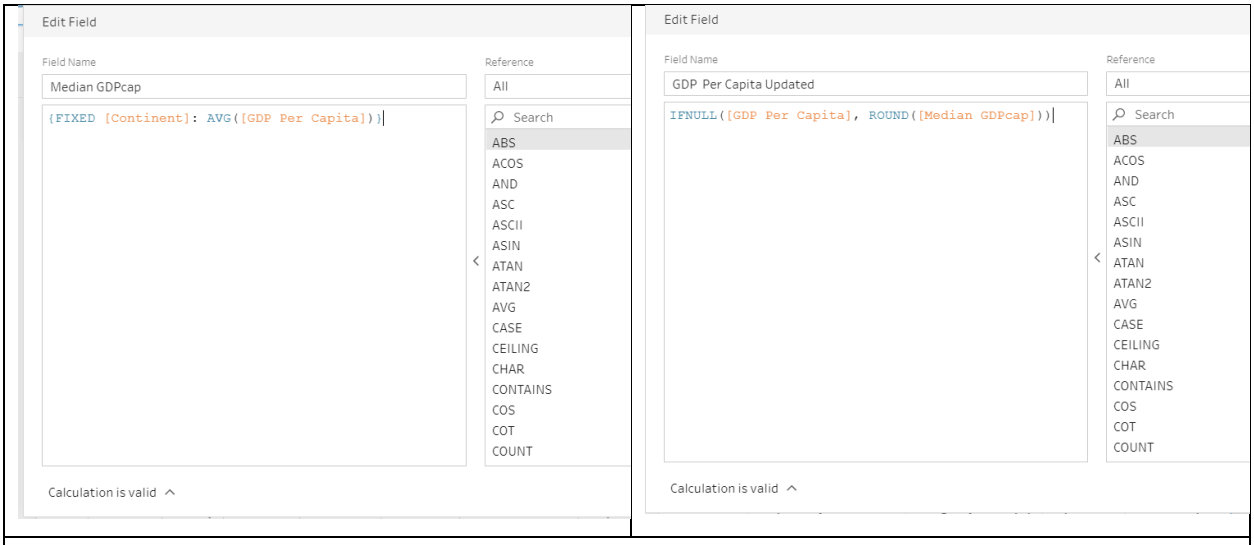


Figure 6.4: Null value in GDP Per Capita and HDI

In the Covid-19 cases dataset, there are null values in GDP per capita and HDI which is not logically right since these figures are always under certain measurement. To solve this issue, the proposed solution is replacing the null with mean value. The mean value is obtained from respective continent the country is in. The reason of doing so is to obtain a complete dataset for further production especially during data modelling.

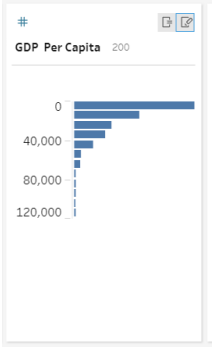
Implementation (Partial)



1. Create **Custom Calculation**
2. Extract A**verage / Median GDP per capital** based on continent using formula:
   `{FIXED [Continent]: AVG([GDP Per Capita])}`
3. **Update** the GDP Per Capita with formula:
   `IFNULL([GDP Per Capita], ROUND([Median GDPcap], 2))`
4. **Remove** Unwanted Column and **rename** the latest column.

Result (Partial)

After



GDP Per Capita Updated	Median GDPcap
33,483.07	33,483.06866889614
33,483.07	33,483.06866889614
33,483.07	33,483.06866889614
33,483.07	33,483.06866889614
33,483.07	33,483.06866889614

Figure 6.5: Round Up the Decimals

The median is being calculated by averaging the GDP Per Capita grouped by the continent. Then, the null GDP Per Capita is being detected and changed to a rounded-up GDP median with two decimals. This is to present the data in a better and cleaner way.

- Additional **Multidimension FIXED** function for mean

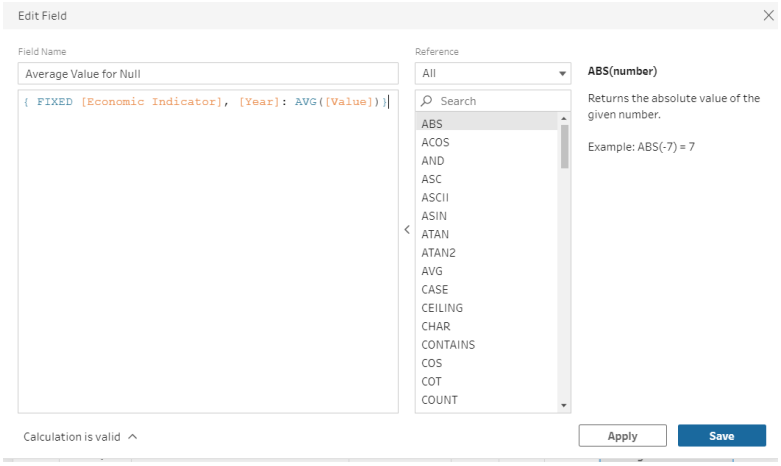


Figure 6.6: Multidimension FIXED

In the Economic Overview Data, there is a need to retrieve the mean / average by referring to multidimension columns. To apply it, the FIXED function will hold two variables for the calculation.

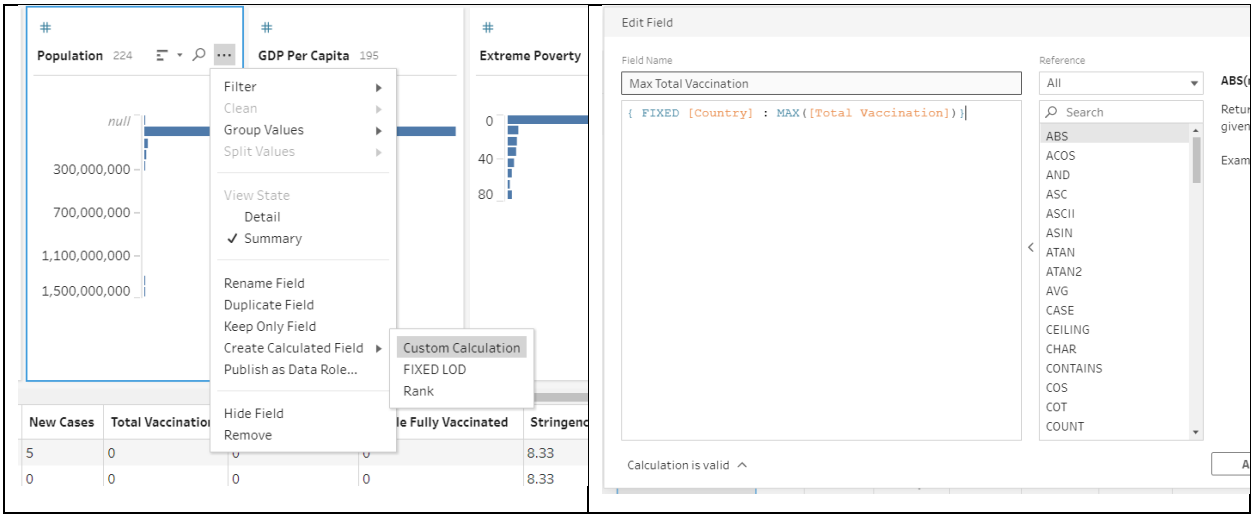
- **Scenario 4:** Replace with appropriate value

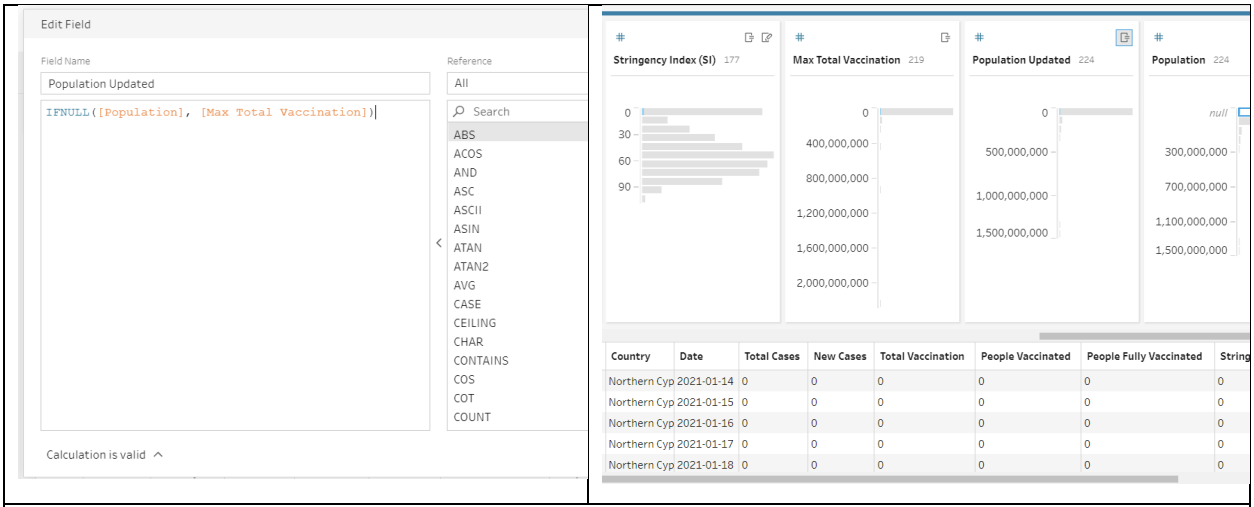
Integer Case

Country	Date	Total Cases	New Cases	Total Vaccination	People Vaccinated	People Fully Vaccinated	Stringency Index (SI)	Population
Northern Cyp	2021-09-22	0	0	0	0	0	0	*null*
Northern Cyp	2021-09-23	0	0	0	0	0	0	*null*
Northern Cyp	2021-09-24	0	0	0	0	0	0	*null*
Northern Cyp	2021-09-25	0	0	0	0	0	0	*null*
Northern Cyp	2021-09-26	0	0	491,206	227,155	223,142	0	*null*

Illogical null value in population is found in Northern Cyprus Country when there are values in total vaccination. Through research, the population counts online are inconsistent and inaccurate. As an alternative, the maximum value in total vaccination (491206) is being considered as the population of the country.

Implementation





1. Create **Custom Calculation**
2. Extract **maximum total vaccination** based on country using formula:

   ```
   { FIXED [Country] : MAX([Total Vaccination])}
   ```

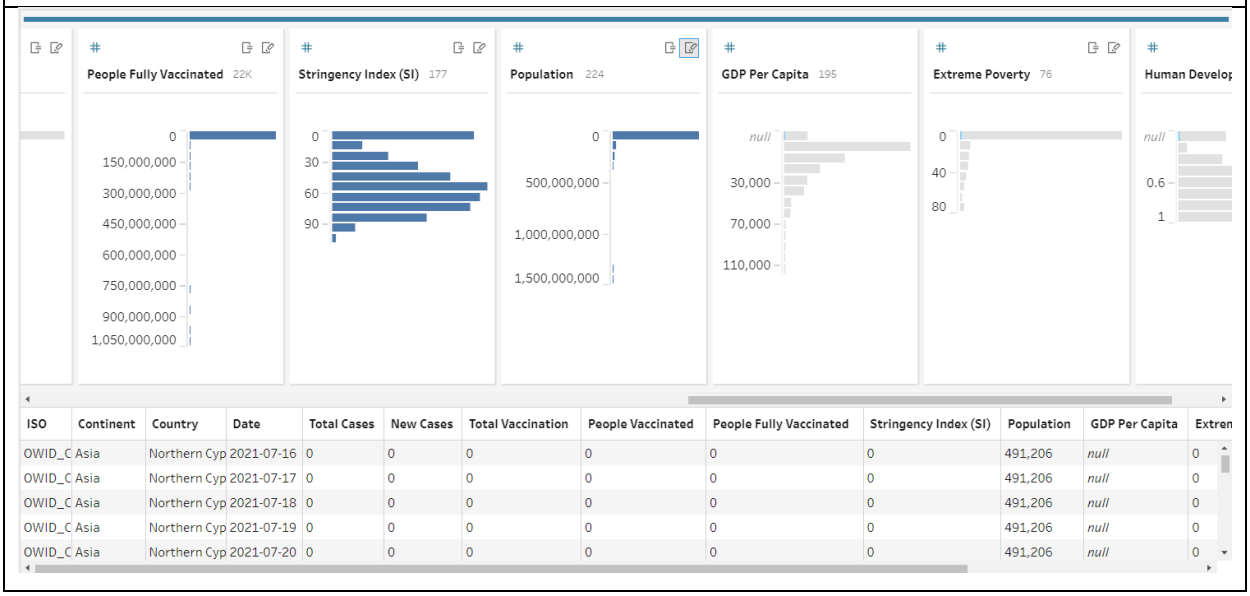
3. **Update** the Population with formula:

   ```
   IFNULL([Population], [Max Total Vaccination])
   ```

4. **Remove** Unwanted Column and **rename** the latest column.

Result

After



The first formula is targeting each unique country value and extracting the max value from total vaccination. The second formula IFNULL() performs as if the population is blank, the data value will be updated with the maximum total vaccination of the country retrieved earlier. Otherwise, no edits are applied to the data column (Tableau, 2021).

String Case

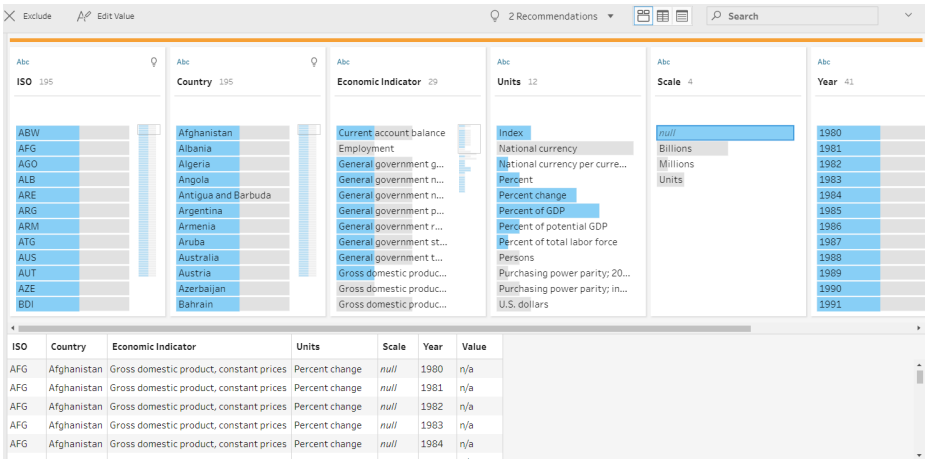
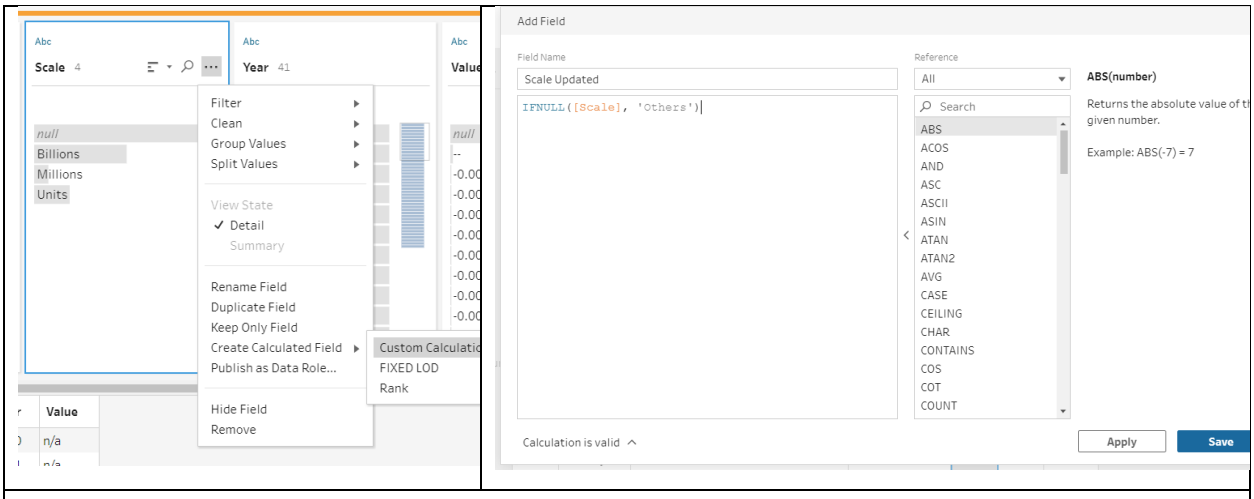


Figure 6.7: Null value in Scale

In the Economic Overview Data, there are null values in the scale where null value is advisable to handle before any further steps on data research. Hence, to solve this problem, the null values are being proposed to replace with “Others” instead of leaving it as null which is also more meaningful.

Implementation (Partial)

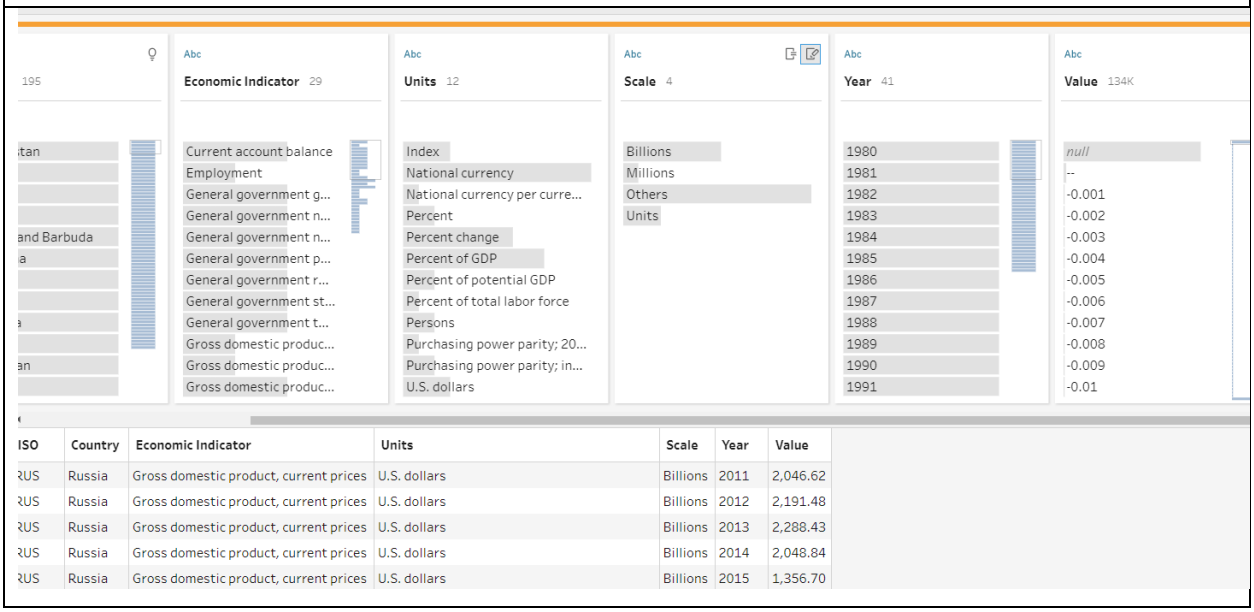


1. Create **Custom Calculation**
2. **Update** the Scale with formula:

   `IFNULL([Scale], 'Others')`
3. **Remove** Unwanted Column and **rename** the latest column.

Result

After



Task 4: Edit Value (Special Case)

- Meaningful Terms

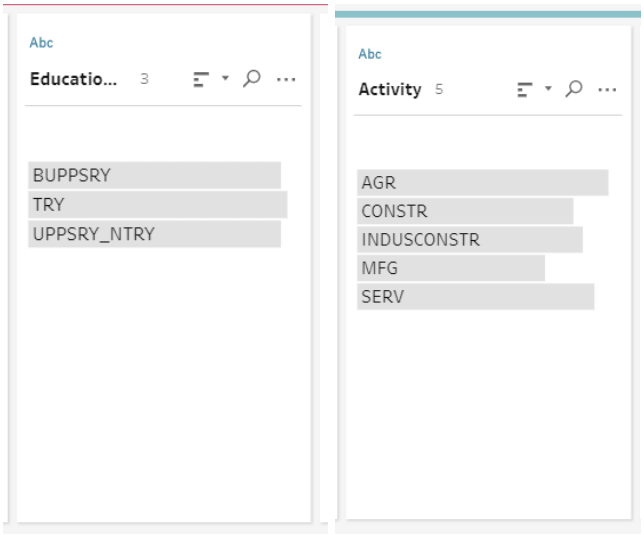


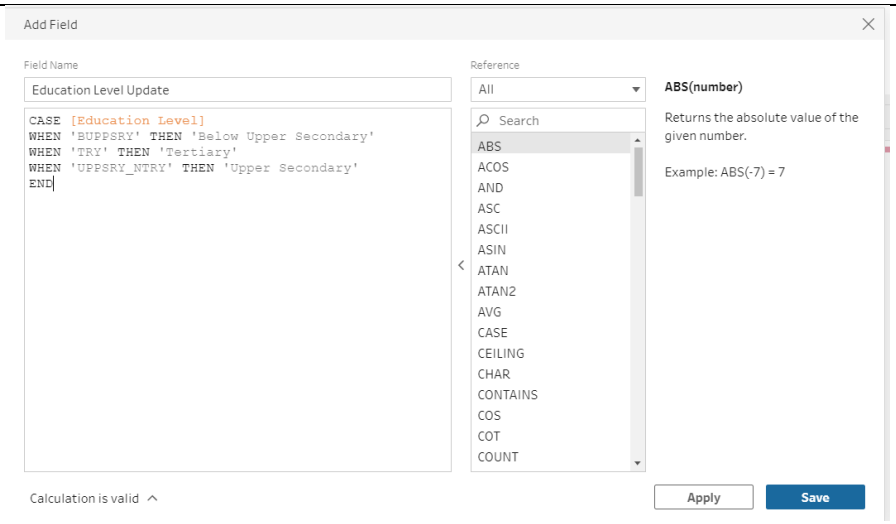
Figure 6.8: Unexplainable Column Data

In the employment and unemployment data, the description of certain grouping is not clear when researcher look through it without studying any reference. For example, the education level and activity columns store data value that is presented in short form. Therefore, some amendment on the data value should be implemented to ensure the data quality and easier for the researcher to understand the data.

Value references:
https://data.oecd.org/emp/employment-by-education-level.htm#indicator-chart
https://data.oecd.org/unemp/unemployment-rates-by-education-level.htm#indicator-chart

The implementation should avoid any manual changes on the value where it is being implemented through the system automation.

Implementation (Partial)



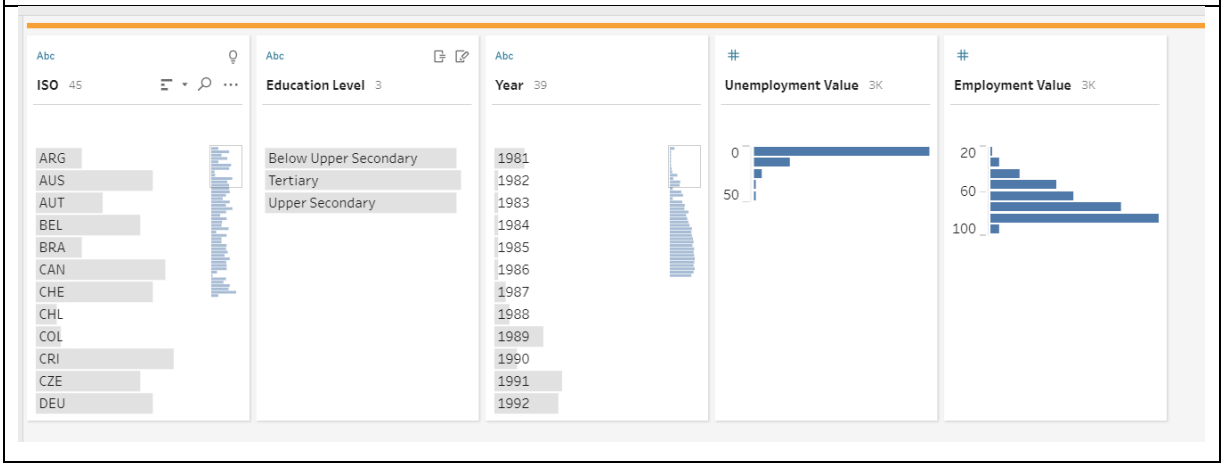
1. Create **Custom Calculation**
2. **Update** the Education level with formula:

```
CASE [Education Level]
WHEN 'BUPPSRY' THEN 'Below Upper Secondary'
WHEN 'TRY' THEN 'Tertiary'
WHEN 'UPPSRY_NTRY' THEN 'Upper Secondary'
END
```

3. **Remove** Unwanted Column and **rename** the latest column.

Result (Partial)

After



- Change Term (Standard)

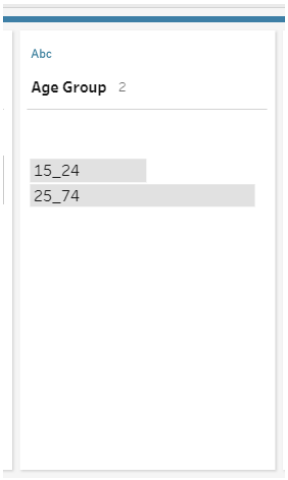
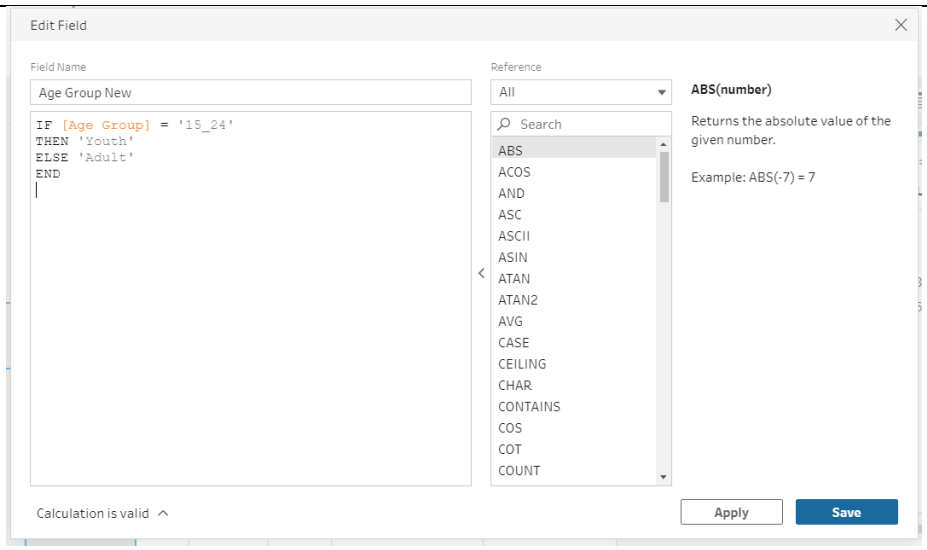


Figure 6.9: Age Group Terminology

In the age group, it is being displayed as 15_24 and 25_74 category. However, it is not descriptive which should be replace with better and standard terminology. The swappings are as listed:

- 15_24: Youth
- 25_74: Adult

Implementation



1. Create **Custom Calculation**

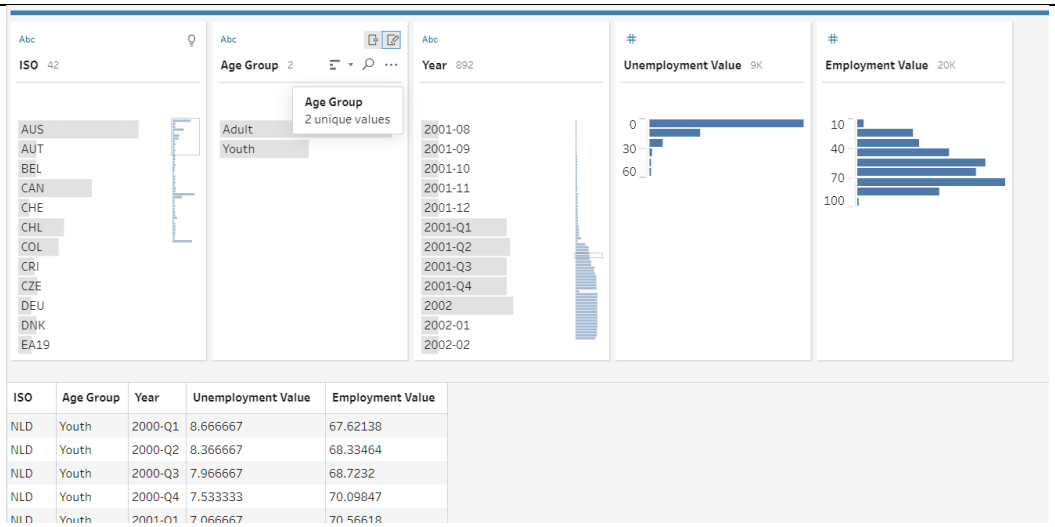
2. **Update** the Education level with formula:

```
IF [Age Group] = '15_24'
THEN 'Youth'
ELSE 'Adult'
END
```

3. **Remove** Unwanted Column and **rename** the latest column.

Result

After



- Absolute integer

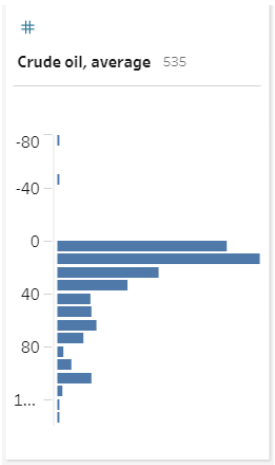
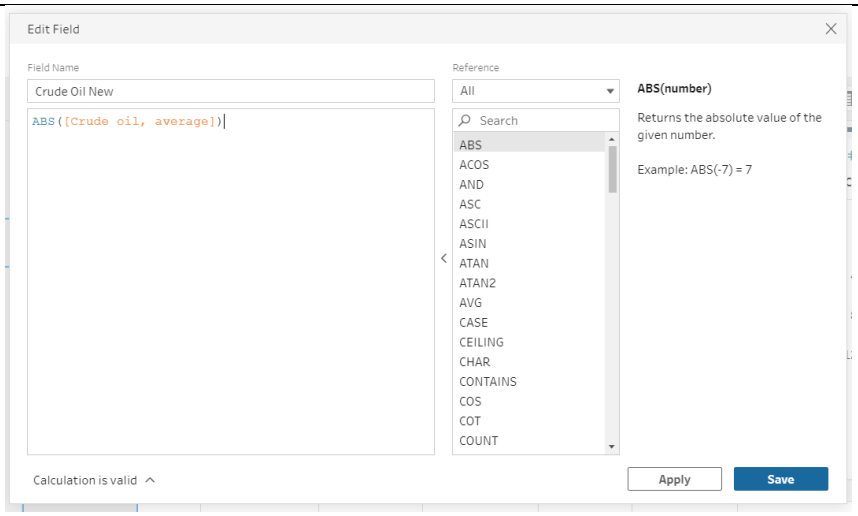


Figure 6.10: Incorrect Negative Value in Commodity Price

Logically, the crude oil commodity price should not be negative as it is a positive price. It is because it is not a financial change but a static pricing. So, the value supposed to be fixed to positive.

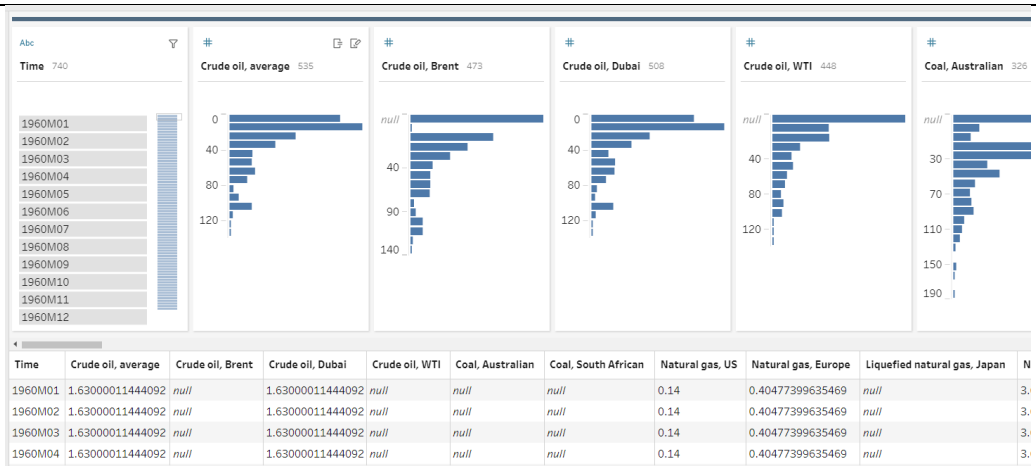
Implementation



1. Create **Custom Calculation**
2. **Update** the Crude Oil, average with formula:
   `ABS([Crude oil, average])`
3. **Remove** Unwanted Column and **rename** the latest column.

Result

After



Time	Crude oil, average	Crude oil, Brent	Crude oil, Dubai	Crude oil, WTI	Coal, Australian	Coal, South African	Natural gas, US	Natural gas, Europe	Liquefied natural gas, Japan
1960M01	1.63000011444092	null	1.63000011444092	null	null	null	0.14	0.40477399635469	null
1960M02	1.63000011444092	null	1.63000011444092	null	null	null	0.14	0.40477399635469	null
1960M03	1.63000011444092	null	1.63000011444092	null	null	null	0.14	0.40477399635469	null
1960M04	1.63000011444092	null	1.63000011444092	null	null	null	0.14	0.40477399635469	null

ABS() function is an absolute mathematical function that transform negative value to positive value where it is implied on the crude oil price to obtain all positive data/

- Remove irrelevant data

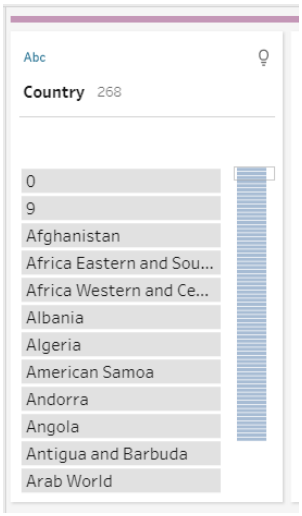
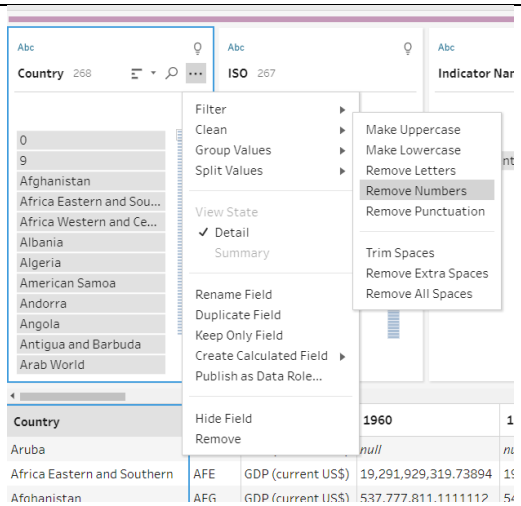


Figure 6.11: Numeric Value in Country Column

In the GDP Value data table, there are numeric values in the Country column where they should be filtered out as it is inaccurate.

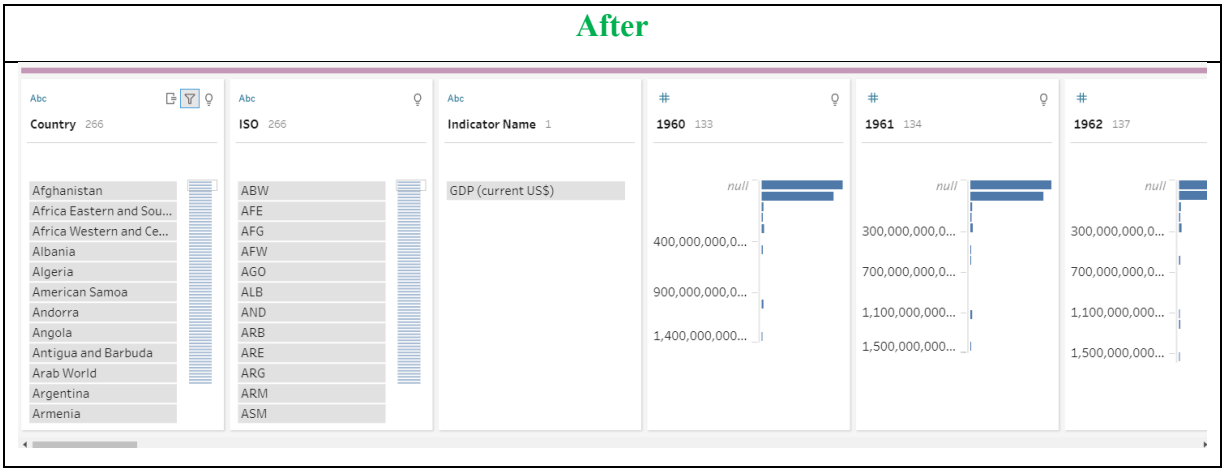
Implementation



1. Select **Remove Numbers**

2. **Note:** Based on observation, the row with number is all null
3. **Proceed** to remove the whole rows

Result



Note:

- Pivot column to row and null value is performed later

- Standardized integer data values

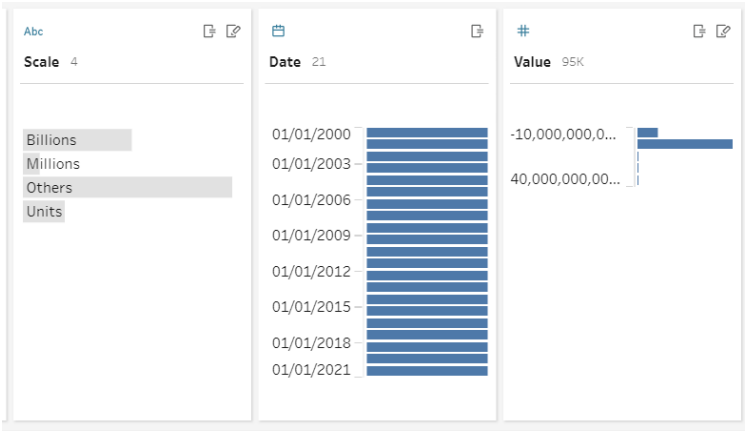
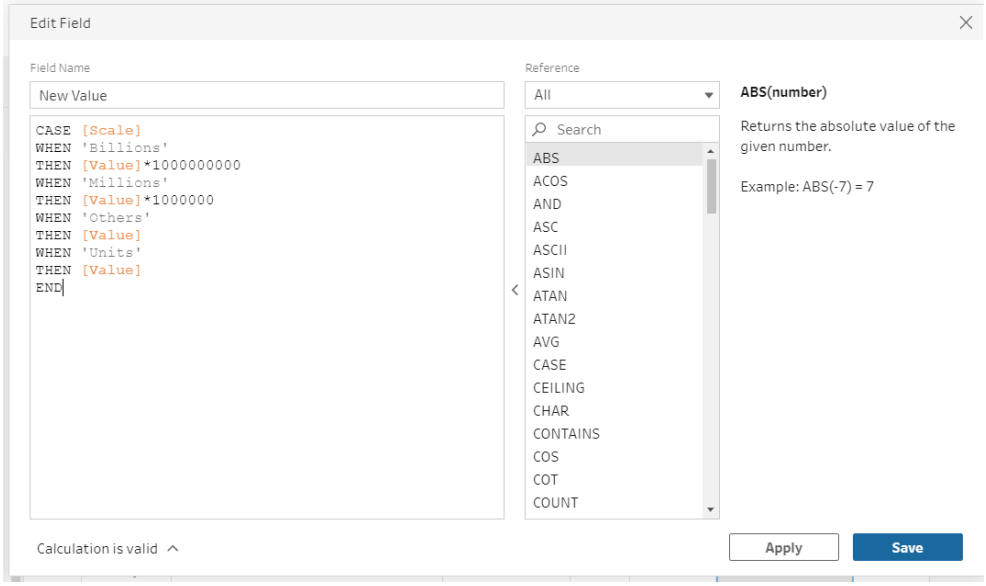


Figure 6.12:Unnecessary dimensional data columns

In the economic overview data, there are unnecessary integer value matching with respective scale. To view the data more efficiently, it is being suggest that the Billions and Millions should directly applied to the numbers, meanwhile for Units and Others scales, the data value remain as it is.

Implementation

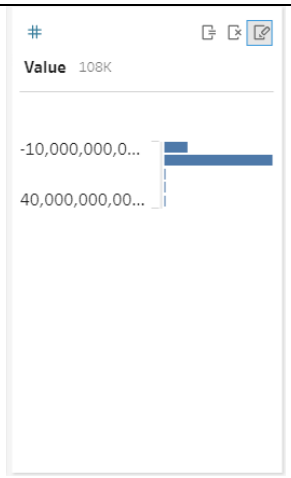


1. Create **Custom Calculation**
2. **Update** New Value with formula:

```
CASE [Scale]
WHEN 'Billions'
THEN [Value]*1000000000
WHEN 'Millions'
THEN [Value]*1000000
WHEN 'Others'
THEN [Value]
WHEN 'Units'
THEN [Value]
END
```

3. **Remove** Unwanted Column and **rename** the latest column.

Result

After


Task 5: Split Data

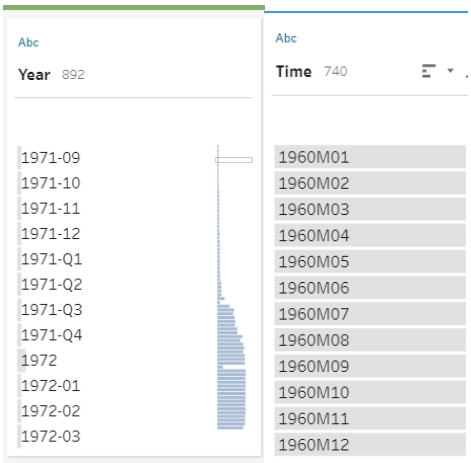
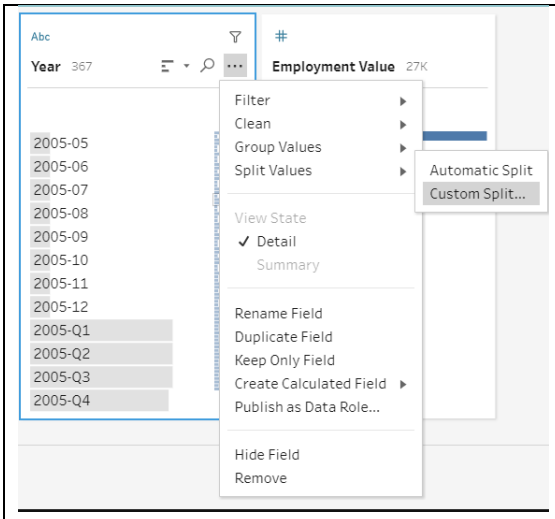


Figure 6.13: Unformatted Date

Based on the date related data in the image, the date is inconsistent where the year is followed by either month or quartile. To solve it, data splitting should be conducted so researcher can obtain a clean year data.

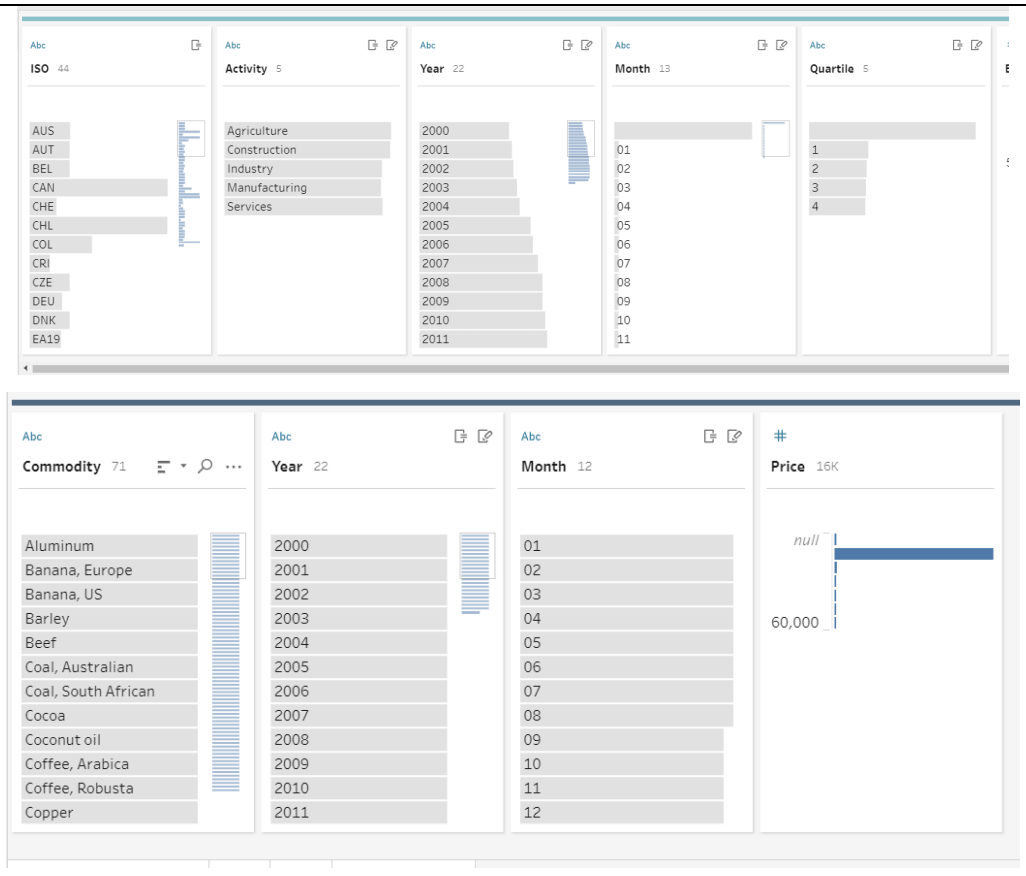
Implementation



1. Select **Split Values > Custom Split**
2. Set **separator.** (eg. -, M, Q)
3. Set **field number** according to need (eg. 2)
4. **Rename** Column and **remove** unwanted columns

Result

After



Task 6: Swap Case

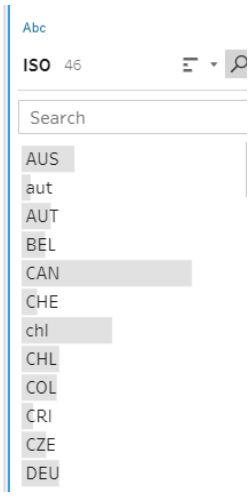
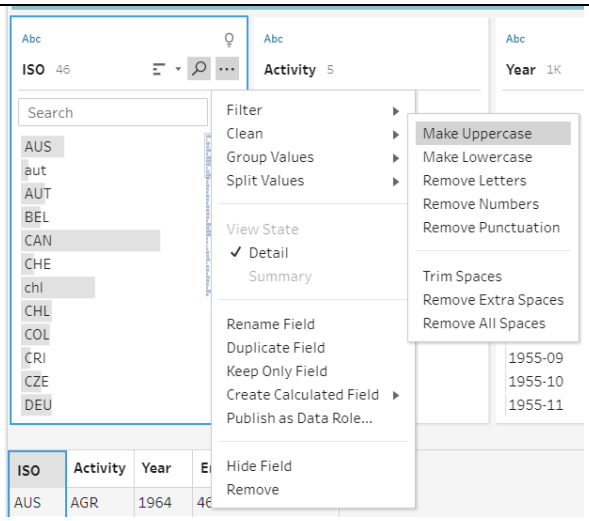


Figure 6.14: Inconsistent Casing in ISO Column

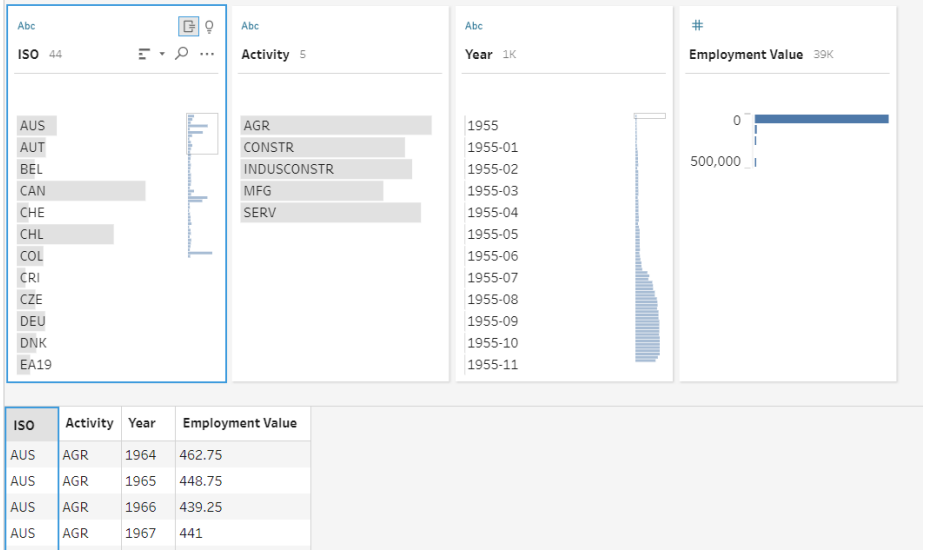
In the ISO column in the figure above, there are lowercase and uppercase versions where they should be standardized to uppercase since it is a ISO country code.



1. Select **Make Uppercase**

Result

After



Task 7: Transpose / Pivot

- Columns to Rows

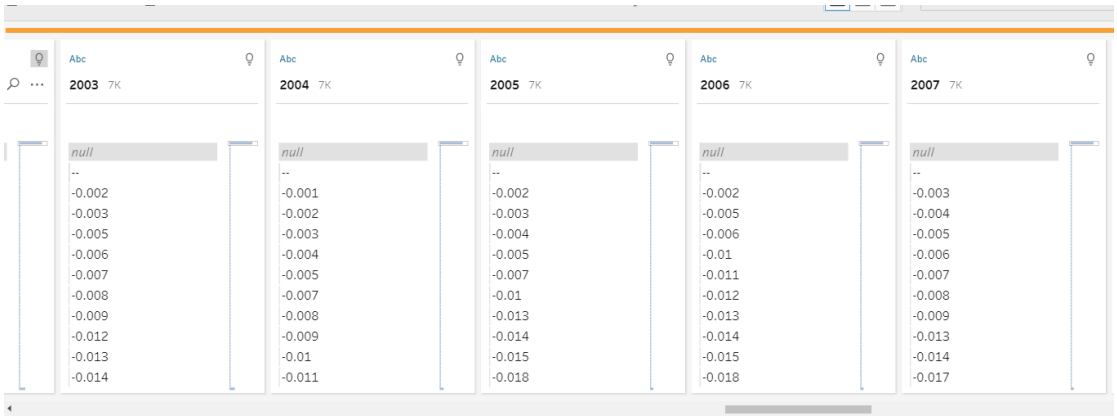
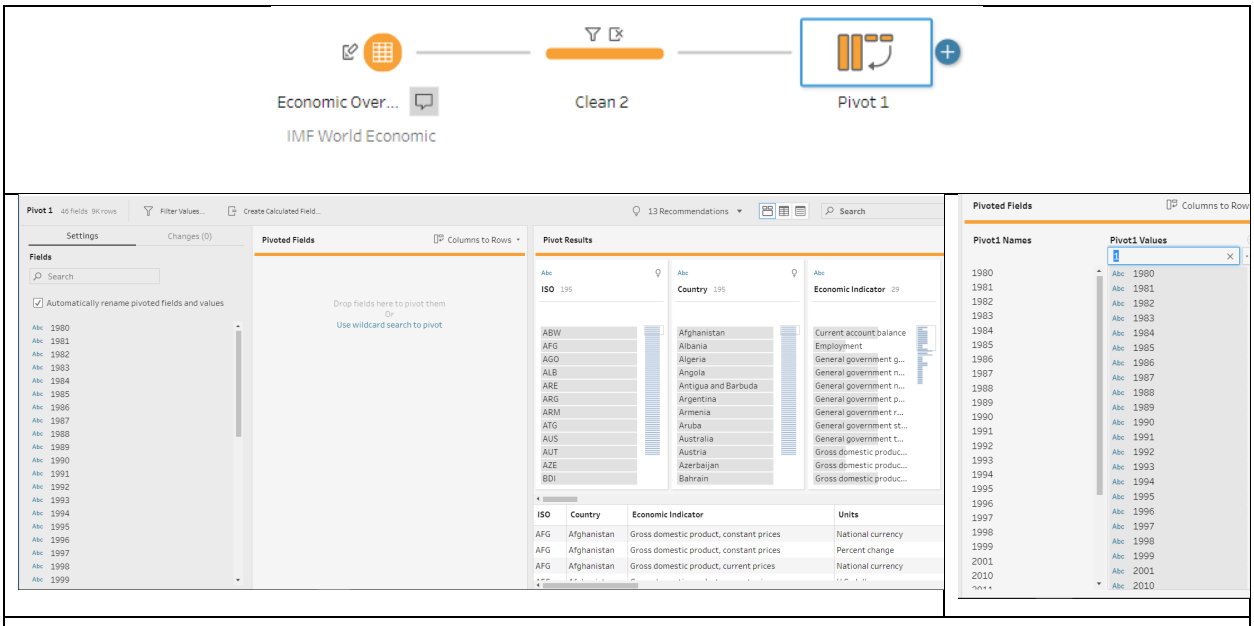


Figure 6.15: Year Data Displayed as Column

In the messy data, it is recognized that there is a need to transpose the data through columns to rows. For example, in the screenshot above, the year data supposing to be part of the row data record. The reason of transposing it to a row is to create an easier navigation between data and polish the dimensional data representation.

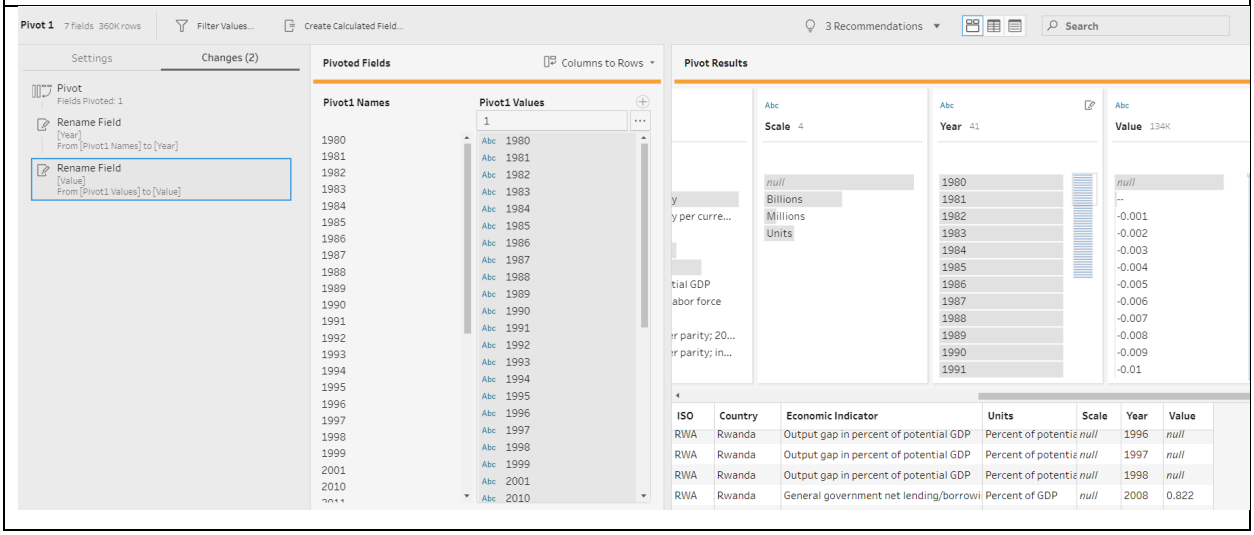
Implementation (Partial)



1. Add **Pivot** Step
2. Use **Wildcard** to automate most of the addition of **Columns to Rows**
3. **Drag** unselected year columns to selection in Pivoted Field.
4. Rename **Pivoted** Column Name

Result (Partial)

After



Note:

- Null values are handled after Pivot steps

- Rows to Columns

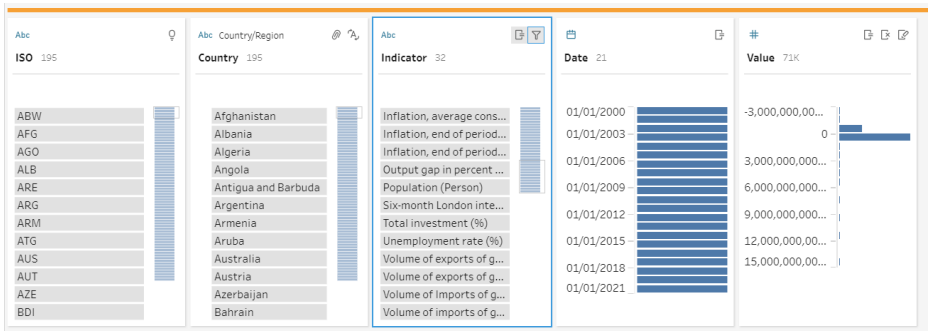
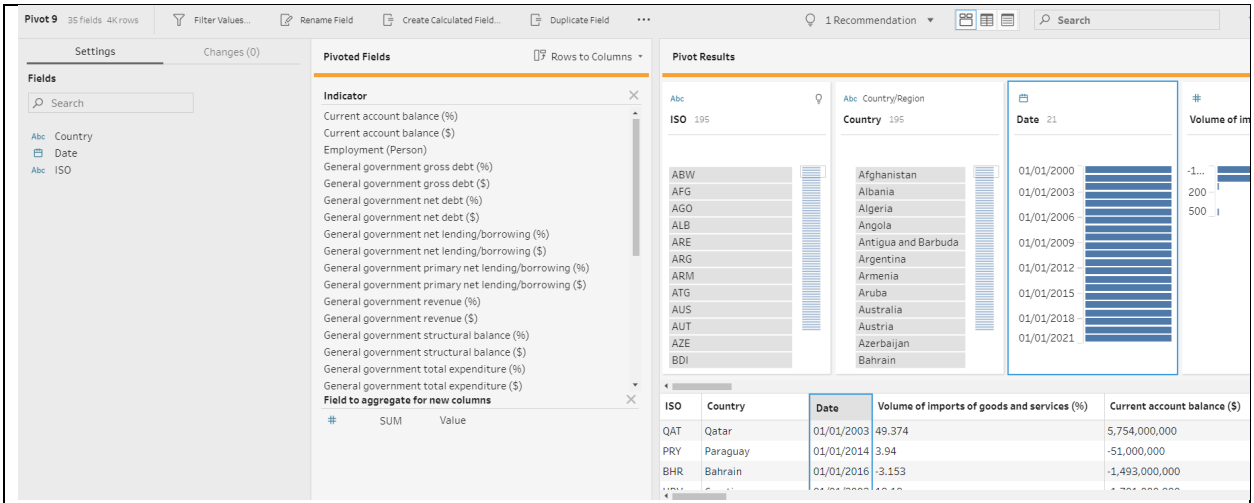


Figure 6.16: Indicators Column

In the Indicators column, there are various type of economic indicators which suppose to split into several columns for clearer analysis especially during data visualization and distribution.

Implementation



1. Add **Pivot** Step
2. Drag Indicator Column to **Pivoted Fields** act as column name / header.
3. Drag Value to **Field to aggregate for new columns**.

Result

After

ISO	Country	Date	Volume of imports of goods and servi...	Current account bal
QAT	Qatar	01/01/2003	49.374	5,754,000,000
PRY	Paraguay	01/01/2014	3.94	-51,000,000
BHR	Bahrain	01/01/2016	-3.153	-1,493,000,000
HRV	Croatia	01/01/2002	18.18	-1,701,000,000
MLI	Mali	01/01/2007	1.303	-454,000,000
FJI	Fiji	01/01/2002	6.27	-20,000,000
EST	Estonia	01/01/2006	20.703	-2,545,000,000
ROU	Romania	01/01/2002	10.671	-1,231,000,000
UZB	Uzbekistan	01/01/2013	8.825	1,631,000,000
HND	Honduras	01/01/2006	5.436	-404,000,000
PHL	Philippines	01/01/2003	5.768	285,000,000
CMR	Cameroon	01/01/2018	2.059	-1,409,000,000
SSD	South Sudan	01/01/2006	9.89	800,000,000
FSM	Micronesia	01/01/2006	9.89	-39,000,000
SVK	Slovak Republic	01/01/2011	7.817	-4,867,000,000
UGA	Uganda	01/01/2020	3.48	-3,430,000,000
EGY	Egypt	01/01/2009	-1.61	-4,424,000,000
TUN	Tunisia	01/01/2011	-3.778	-3.869.000.000

Task 8: Flatten data

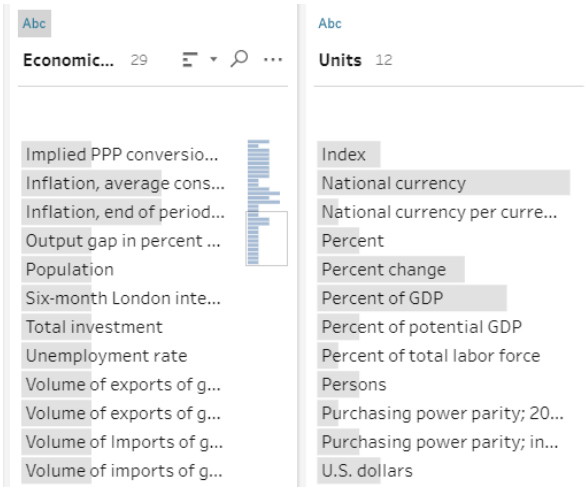
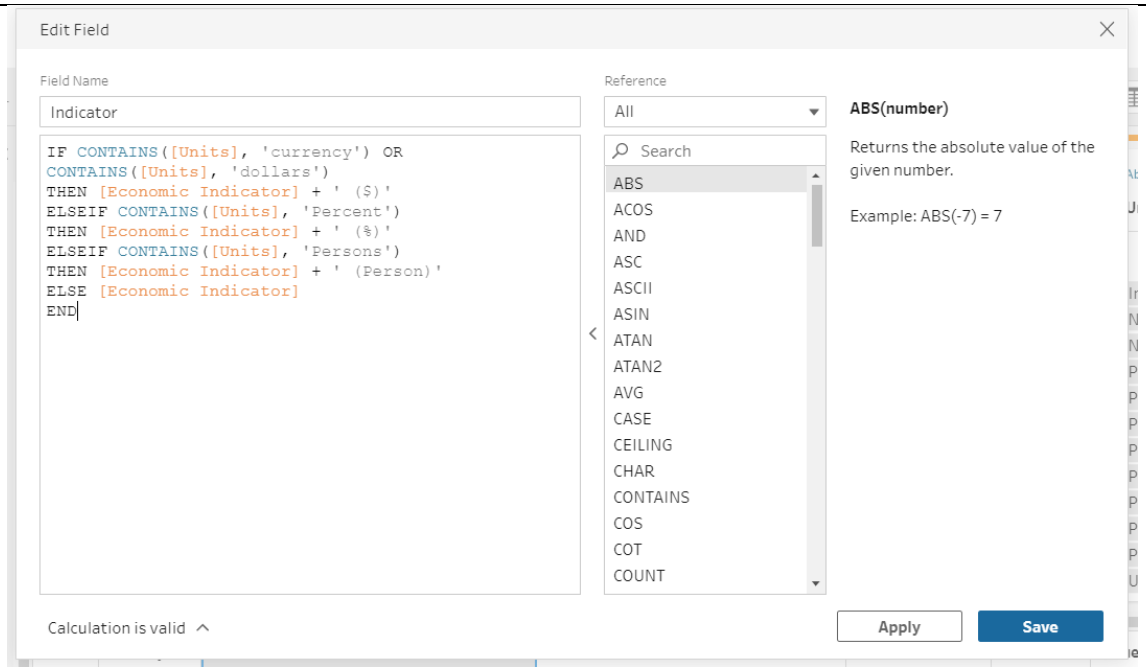


Figure 6.17: Indicators and Unit Name-value_type Pairs

It is being identified that in economic overview data, there are name-value_type pairs which can be flatten into one value. This eventually means that a combination on the economic indicator and place value (Units) is preferably to be implement. As a justification, this quickens the data retrieval process and minimize the wastage on data storage & processing speed. Due to these justifications, the failure on *Tableau Desktop* can minimized to its maximum/

Implementation



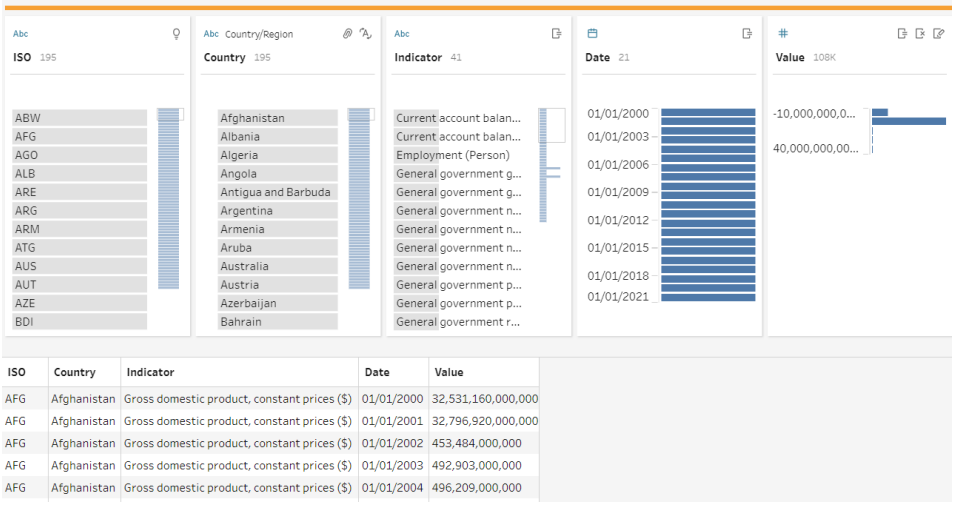
1. Create **Custom Calculation**
2. **Update** Indicator with formula:

```
IF CONTAINS([Units], 'currency') OR CONTAINS([Units], 'dollars')
THEN [Economic Indicator] + ' ($)'
ELSEIF CONTAINS([Units], 'Percent')
THEN [Economic Indicator] + ' (%)'
ELSEIF CONTAINS([Units], 'Persons')
THEN [Economic Indicator] + ' (Person)'
ELSE [Economic Indicator]
END
```

3. **Remove** Unwanted Column and **rename** the latest column.

Result

After



ISO	Country	Indicator	Date	Value
AFG	Afghanistan	Gross domestic product, constant prices ($)	01/01/2000	32,531,160,000,000
AFG	Afghanistan	Gross domestic product, constant prices ($)	01/01/2001	32,796,920,000,000
AFG	Afghanistan	Gross domestic product, constant prices ($)	01/01/2002	453,484,000,000
AFG	Afghanistan	Gross domestic product, constant prices ($)	01/01/2003	492,903,000,000
AFG	Afghanistan	Gross domestic product, constant prices ($)	01/01/2004	496,209,000,000

Note:

- Gross Domestic Products are being filtered since the value are available in other datasets.

The formula explains whenever the Units (value type) contains respective words, the value of Economic Indicators is being amend as listed.

Task 9: Filter Data

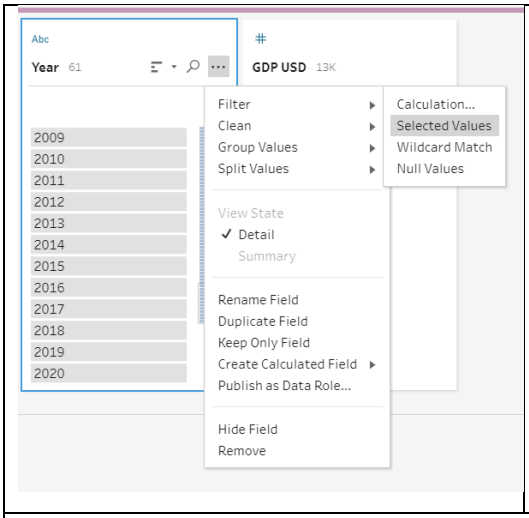
Date chosen: 2000 – 2021

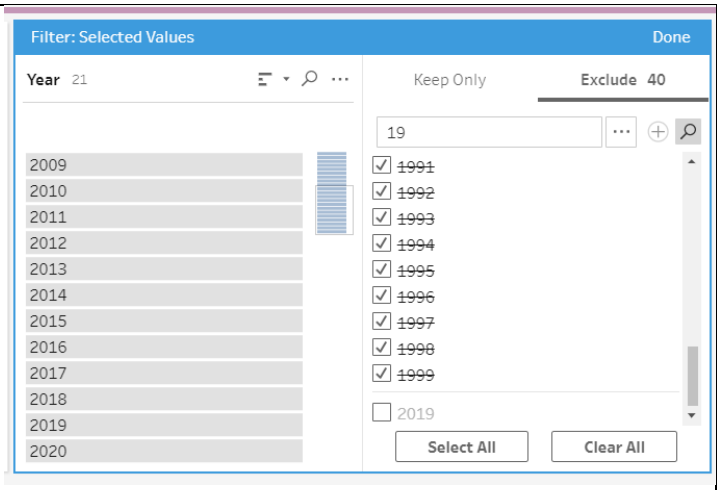
2000 – 2018 - Normal Economic Growth

2019 – 2021 - Covid-19 Specific

The reason of choosing this date timestamp is to compare the normal growth of the economic with the economic changes during Covid-19 period. Secondly, concerning on the performance and efficiency of data retrieval in the platform.

Implementation (Partial)

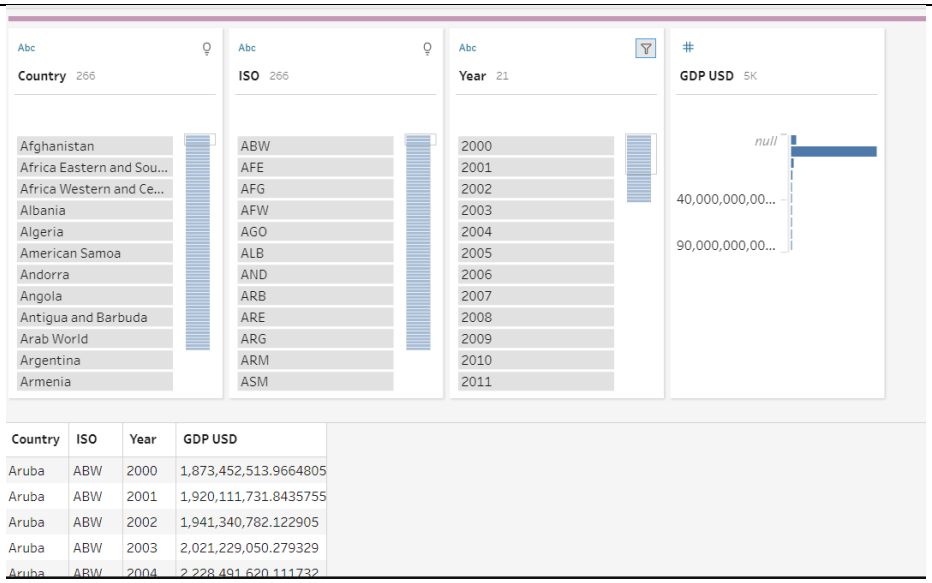




1. Select **Filter > Selected Values**
2. Exclude **19xx** except for 2019 with 19 in it

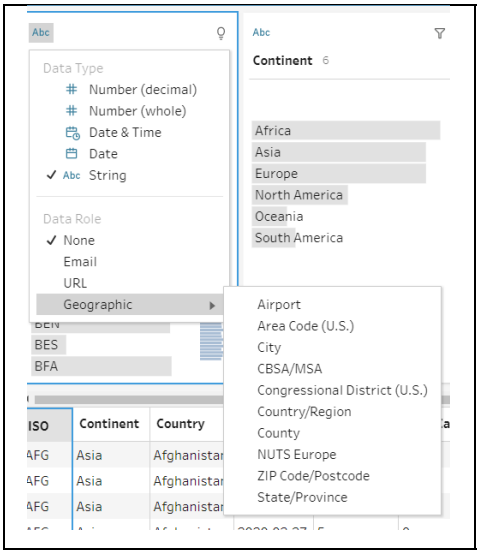
Result

After



Task 10: Data Type Conversion

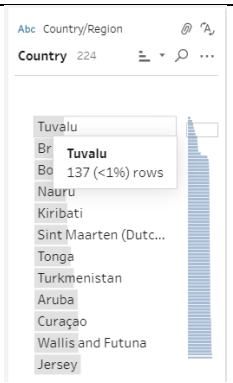
Implementation (Partial)



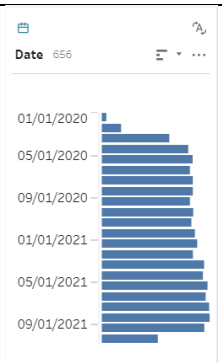
1. Select **Data Type**
2. Select **Data Role** (Optional)

Result

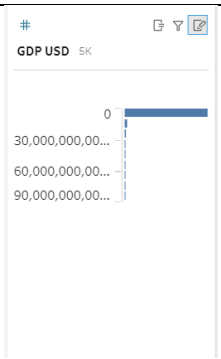
String, **Data Role:** None



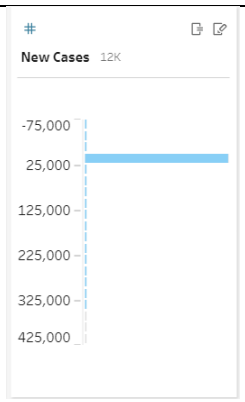
String, **Data Role:** Country/Region



Date

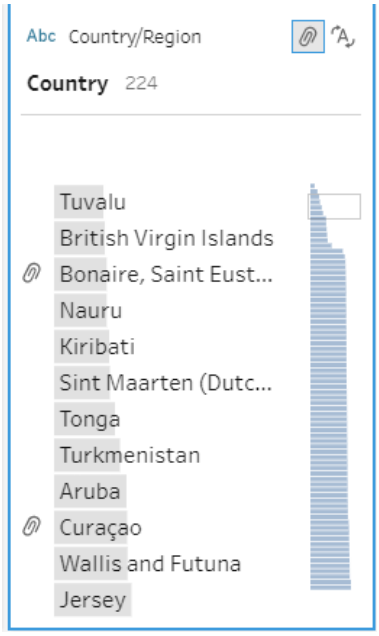


Integer (Decimal)



Integer (Whole)

Additional



There is a need to manual edit some country name to the correct one so that it is noticeable by the Tableau System.

Remarks

- ISO 3166-1 alpha 3 is not identified by the Tableau on Country/Region data role. Might be affected by **System Error**

Task 11: Joining Data

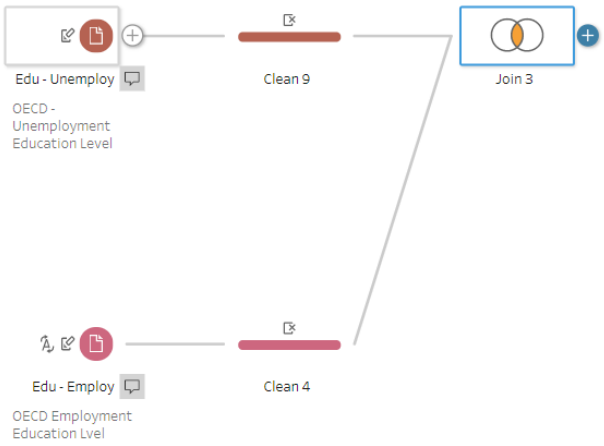
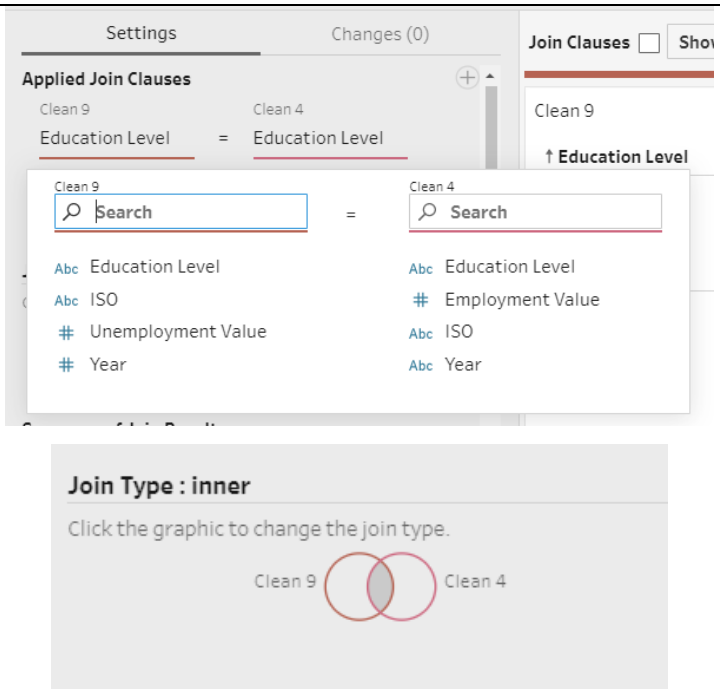


Figure 6.18: Education Level (Unemployment + Employment)

The education level in both employment and unemployment data is similar and it is suitable to join the data. By joining the data, the virtual space use in data mining is preferably small which then improve the performance of data retrieval.

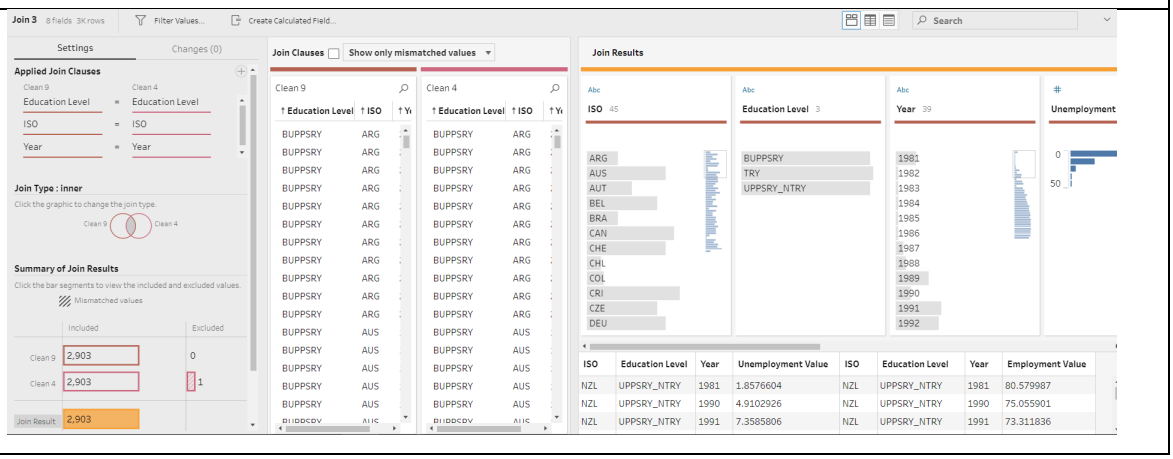
Implementation (Partial)



1. Edit **Join Clauses**
2. Select **Join Type**

Result

After



Note:

- Remove **duplicated columns** in new cleaning step.

Task 12: Date Manipulation

- Year Only Date Creation

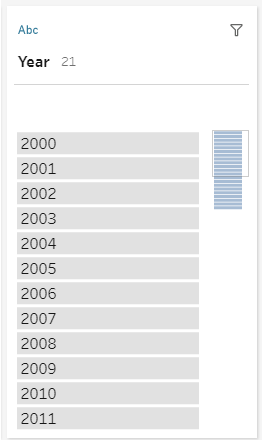
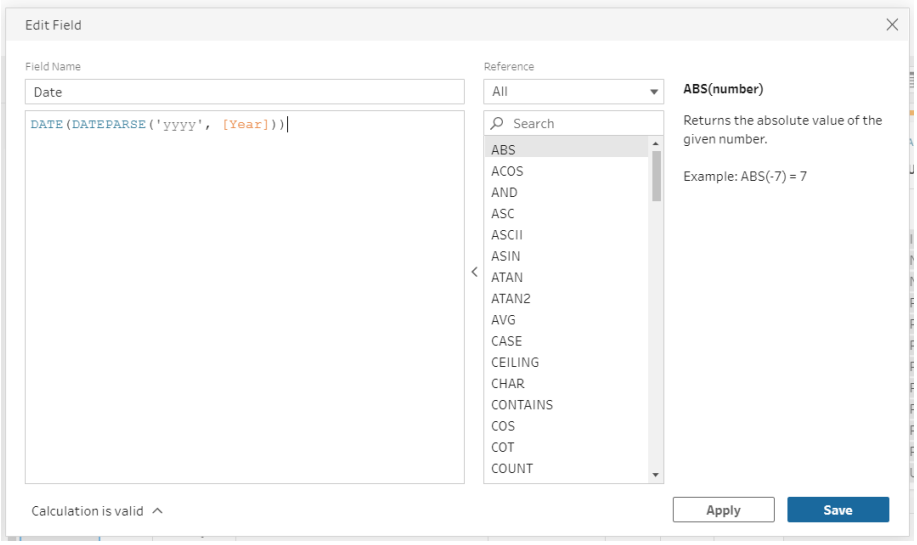


Figure 6.19: Year Data Stored in String Data Type

There are data column that only has year data where it is being stored as String data type. The implementation of switching the column to date data type will cause data distortion on the particular column. Therefore, the case is being resolved through *Tableau* built-in functions.

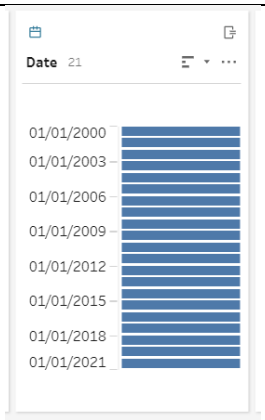
Implementation (Partial)



1. Create **Custom Calculation**
2. **Create** new field – Date with formula:
   `DATE(DATEPARSE('yyyy', [Year]))`
3. **Remove** Unwanted Column and **rename** the latest column.

Result

After



The DATEPARSE function selects the assigned YEAR column's value as 'yyyy' format and form a brand-new date time. The date time is then being through DATE function where only date is remain as the end value (Tableau, 2021).

- Complete Date Creation

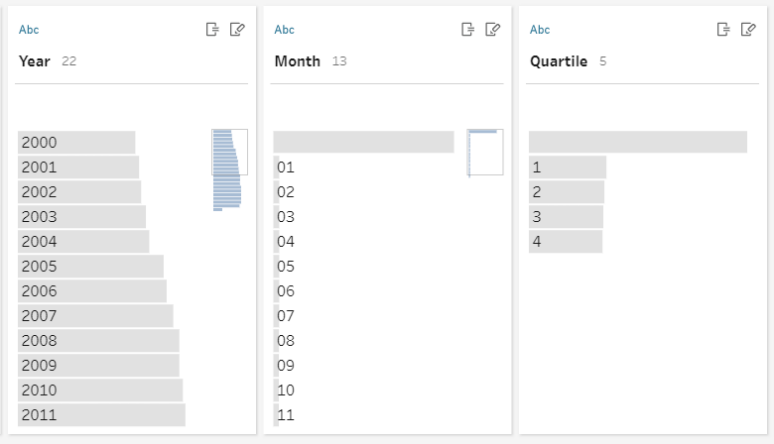
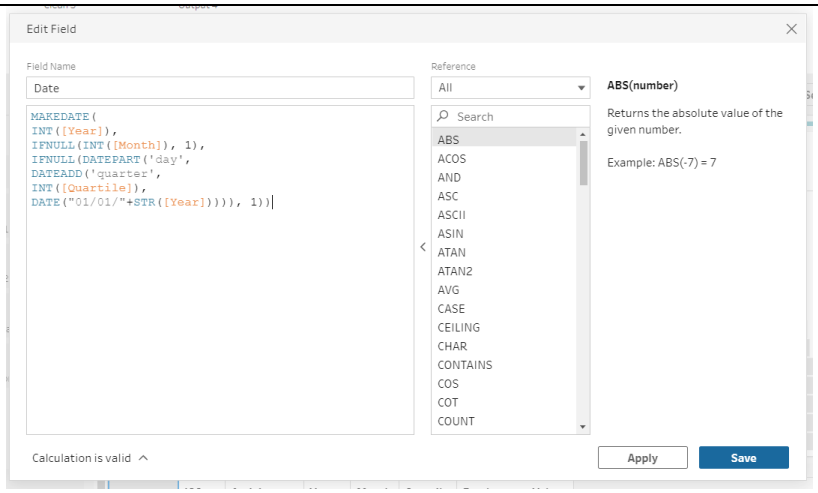


Figure 6.20: Existing Year, Month & Quarter Columns

In the diagram above, the data has existing year, month and quarter data but they are all in String data type. As mentioned above, direct swapping of data type will cause data distortion so in this case built-in functions are being implemented again.

Implementation (Partial)



1. Create **Custom Calculation**

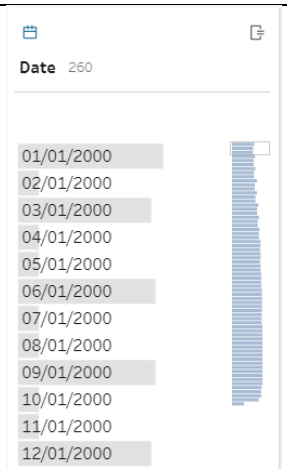
2. **Create** new field – Date with formula:

```
MAKEDATE(
INT([Year]),
IFNULL(INT([Month]),
IFNULL(INT(INT([Quartile])/4 * 12),1)),
1)
```

3. **Remove** Unwanted Column and **rename** the latest column.

Result

After



To explain the formula, generally the function will create a date according to the year, month and day (constant one). Whenever the month or quarter is empty or null, the month will be filled in with one. For case where month data is not ready, the month will be calculated through quarter data.

For case there are no quarter data, formula used is:

```
MAKEDATE(
INT([Year]),
IFNULL(INT([Month]),1),
1)
```

Task 13: Filter Error Data

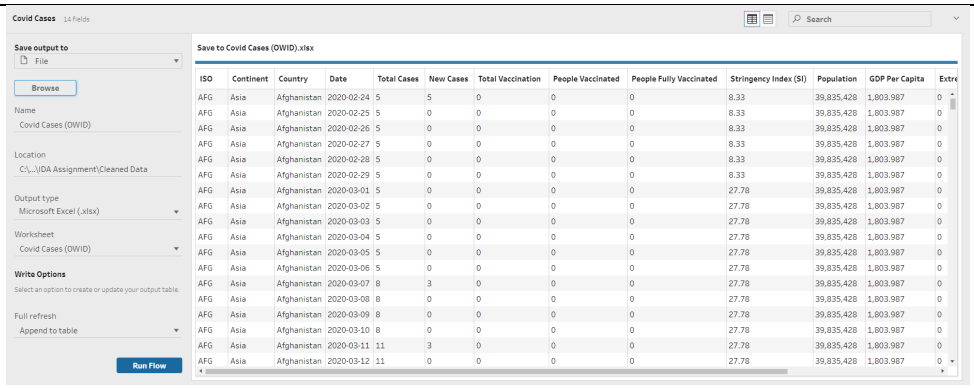
Figure 6.21: ISO Code Filtered Step

Scanning through the data, there are incorrect or aggregated self-defined ISO data in the datasets. Therefore, the records are being filtered out to get a perfect cleaned dataset with error free.

Task 14: Output Data

The final step in data cleaning is to output all data for data mining in *Tableau Desktop*.

Implementation

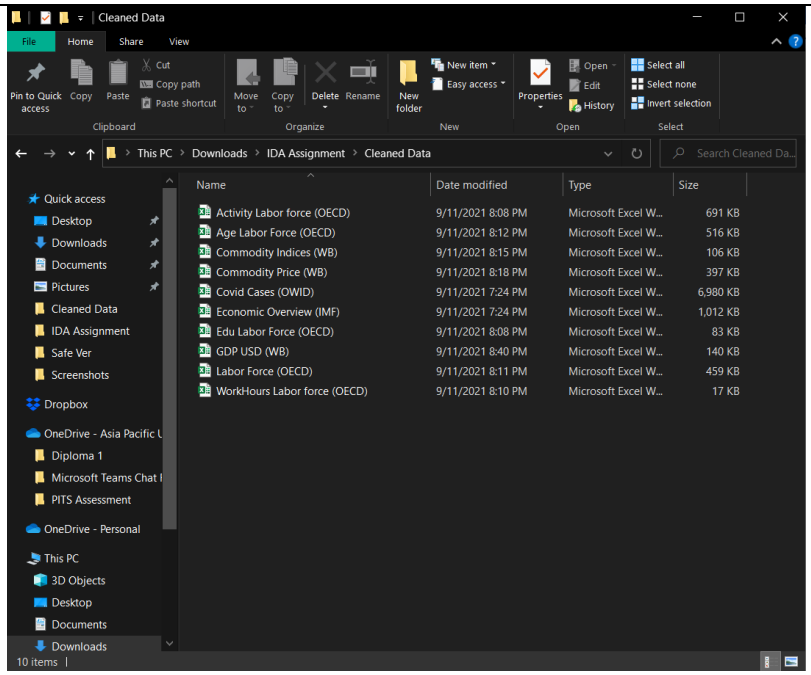


1. **Browse** File Save Location
2. **Name** File
3. **Output Type >** .xlsx
4. Create **Worksheet**

5. **Run flow**

Result

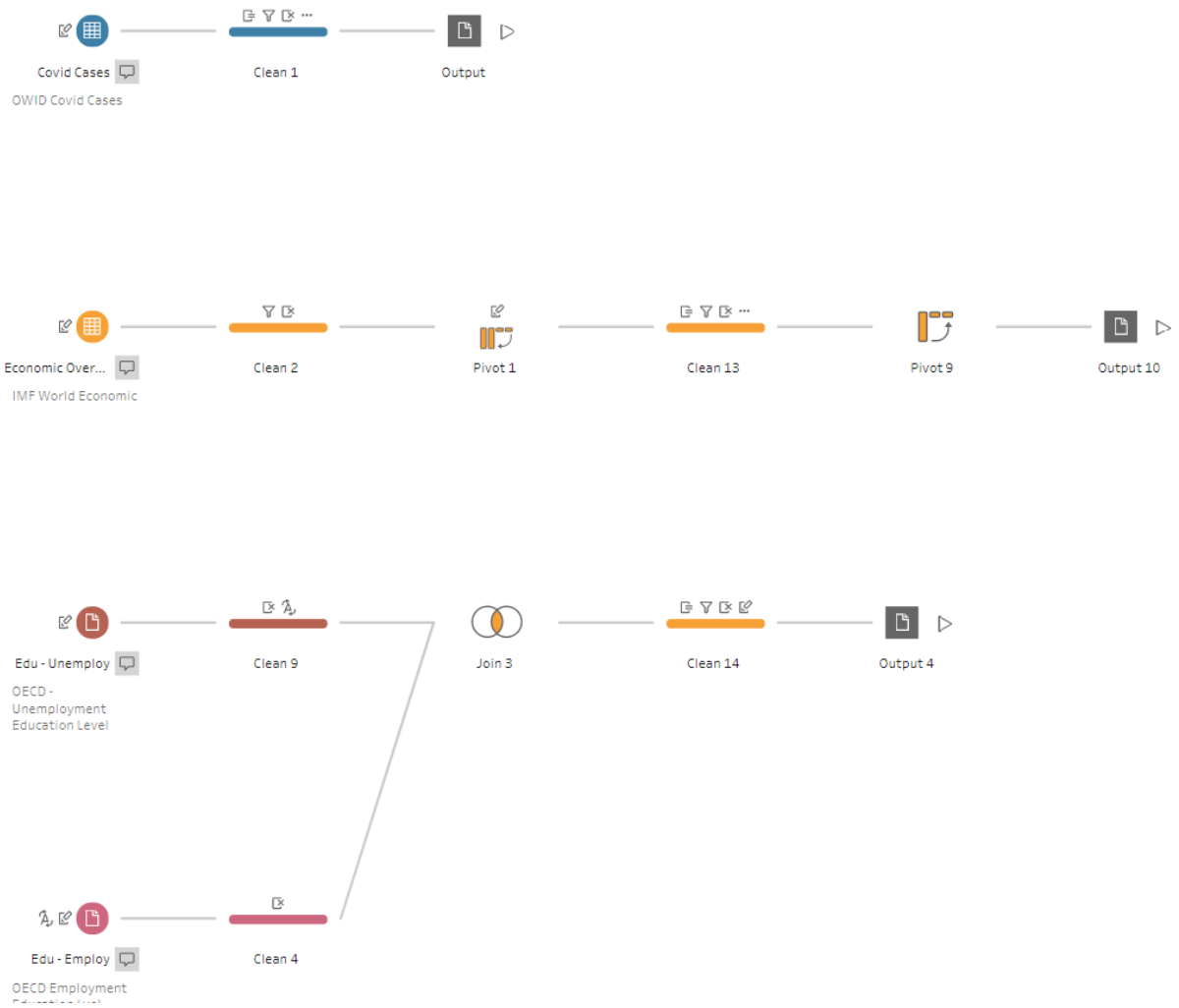
After

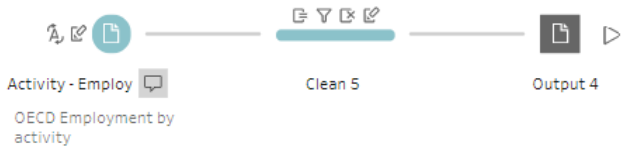


Note:

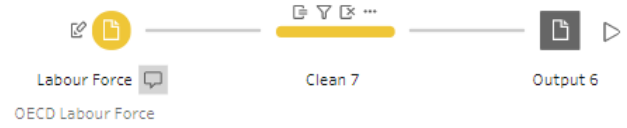
- Each table is saved in respective file for easier recognition rather than compiling all in one.

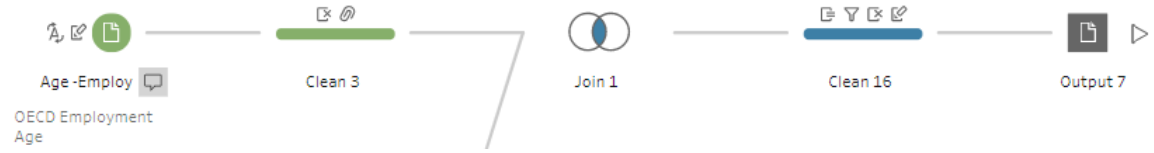
Cleaning Flow

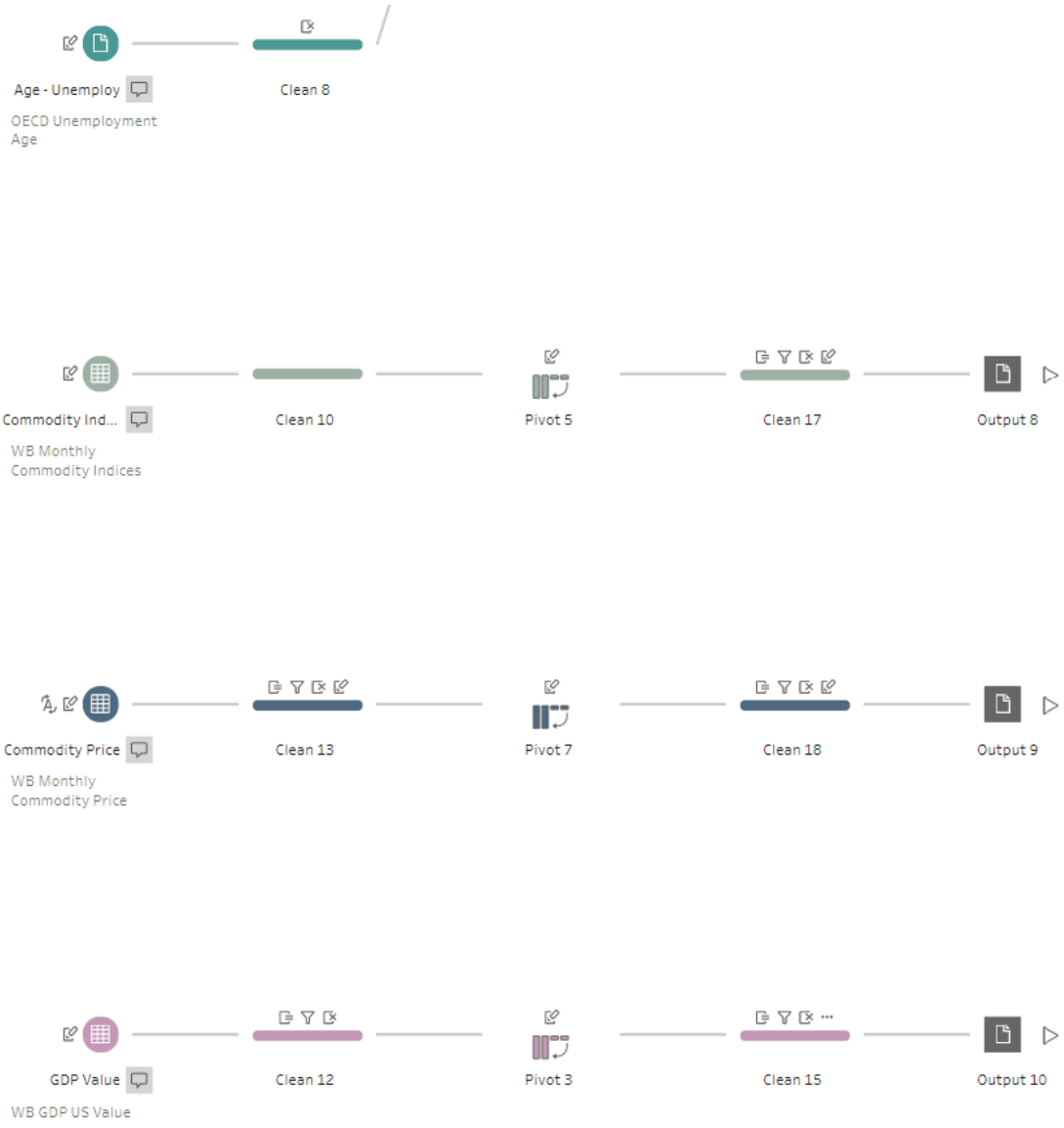



Activity - Employ
OECD Employment by activity
Clean 5
Output 4


Work Hours
OECD Hours Worked
Clean 6
Output 5


Labour Force
OECD Labour Force
Clean 7
Output 6


Age -Employ
OECD Employment Age
Clean 3
Join 1
Clean 16
Output 7



Cleaned data Links:

https://github.com/Laikaiyong/Covid-19-Economic-Impact-Analysis/tree/main/Cleaned%20Data

7.0 Data Mining & Pattern Evaluation

Platform: *Tableau Desktop*

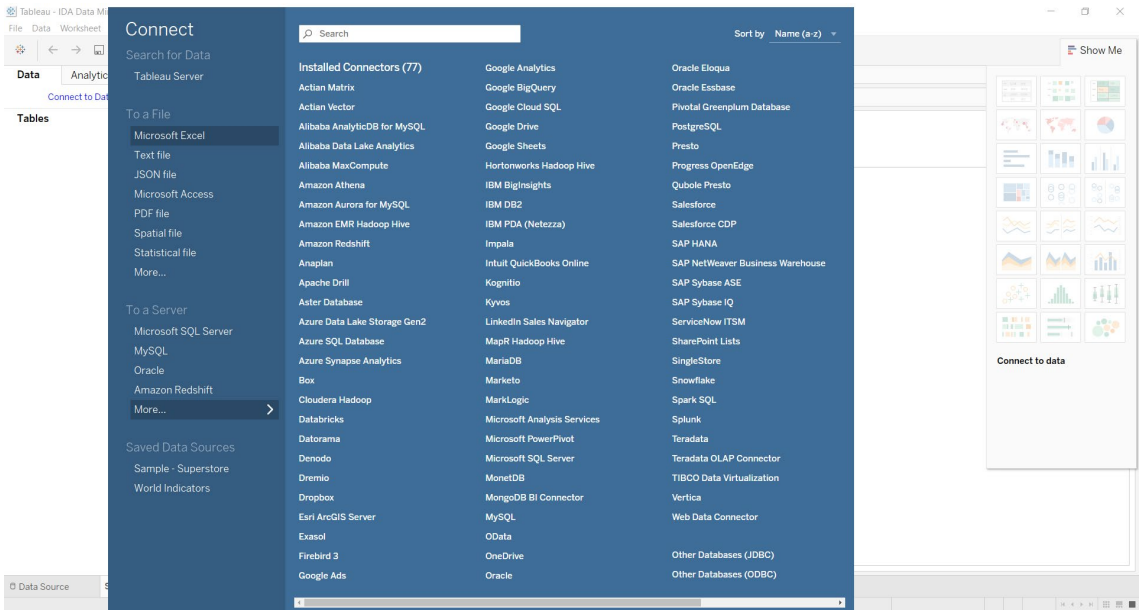


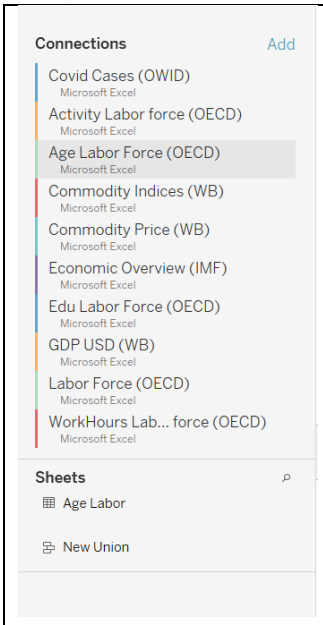
Figure 7.1: Connect to Data Tableau Desktop

After completing the data cleaning, which is the prerequisites for data mining, the cleaned data will be imported to *Tableau Desktop*. The data imported will requires researcher to select the tables required for analysis

7.1 Setup

The setup of a Tableau Project starts with selection of data tables and the establishment on relationship between tables.

Data Tables / Sheets



From the imported Excel spreadsheets, researcher can choose the tables needed for mining. To do so, researchers just need to double-click / drag the table to the connection board.

Relationship

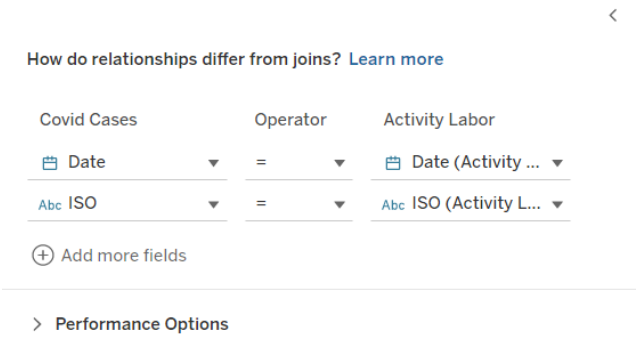


Figure 7.2: Relationship Option

Relationship between tables is between configured and created through common columns. In this research, tables are connected through date, country ISO code and country name. This relationship setting is very important especially in creating an interactive OLAP dashboard.

Figure 7.3: Covid-Countries Economic Relational Schema

Online Published Graphs from this assignment (OLAP capable):
https://public.tableau.com/app/profile/lai.kai.yong.vandyck./viz/IDADataMining/Covid-EconomicIndexOLAPDashboard

7.2 Descriptive analysis

Task 1: OLAP / BI Dashboard

Descriptive analysis is being performed in data mining where the deliverable is the Countries Economic Report that provides multiple pages of OLAP / BI dashboard. In each BI dashboards, data distribution is being presented in variety types of statistical graph and visualization. The statistical visualizations display and combine analysis required Key Performance Indicators (KPIs) providing insights toward strategic planning on economy distribution and operational decision-making (Sutner, 2020). Through a clear connection between data sources, the BI dashboards display data metrics in meaningful graphics incorporating data table view within the charts. The dashboards support data filtering, aggregating data with customizable calculations and standalone data description presented through labels and pop-up tooltips. Furthermore, the BI dashboards support OLAP capabilities.

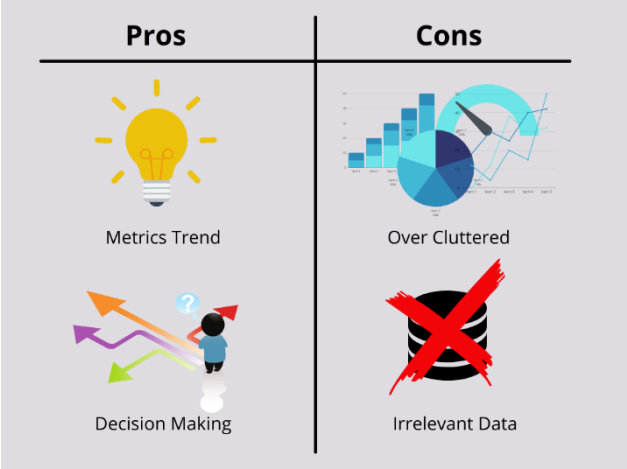


Figure 7.4: Pros and Cons of BI Dashboard

The advantages of building BI dashboard are the ability to discover metrics trend and assist in decision making. The valuable insights obtained from BI dashboard drives better business and operational decision-making. Moreover, multidimensional schema connects huge amount of data where the dashboard simplifies the data into graphics and display the business KPIs in a comprehensible way to the end-users. Overall, BI dashboard ease analysis need and most importantly provide data visualization that help identify data trends. Opposingly, BI dashboard is prone to limitations caused by the way the dashboard is being used especially during the designation of the dashboard. Analysts often distribute too much information to the data visual which over cluttered the dashboard. On the other hand, the selection of data variables will results in data misrepresentation leading to wrong analytics (Sutner, 2020).

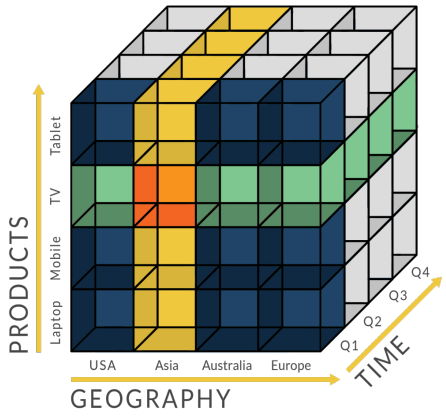


Figure 7.5: OLAP Multidimensional Cube Example
(OLAP.com, 2021)

Meanwhile, OLAP is a technology applied into the BI dashboard which enables researcher to retrieve and query data in different data perspectives. The capability in performing ad hoc data analysis suits data mining research. Generally, OLAP supports the handling of multiple data sources in one warehouse and the presented data cube dispose data dimensions in different view based on users' need. OLAP supports data rolling-up, drill-down, data slicing, data dicing and data pivoting (Wolters Kluwer, 2021). The benefits of implementing OLAP are the high rate of successful decisions. Statistically, the statement is being proved to be true since the more the data acquired for analysis, the more accurate the analysis result will be. Through OLAP, data can be accessed quickly, and researcher can obtain many useful information through the multidimensional data perspectives. Additionally, OLAP provides researcher with the capability to perform sophisticated and effective analysis by manipulating the data in different view along with multidimensional metrics (OLAP.com, 2021).

BI Analysis

- Fully OLAP

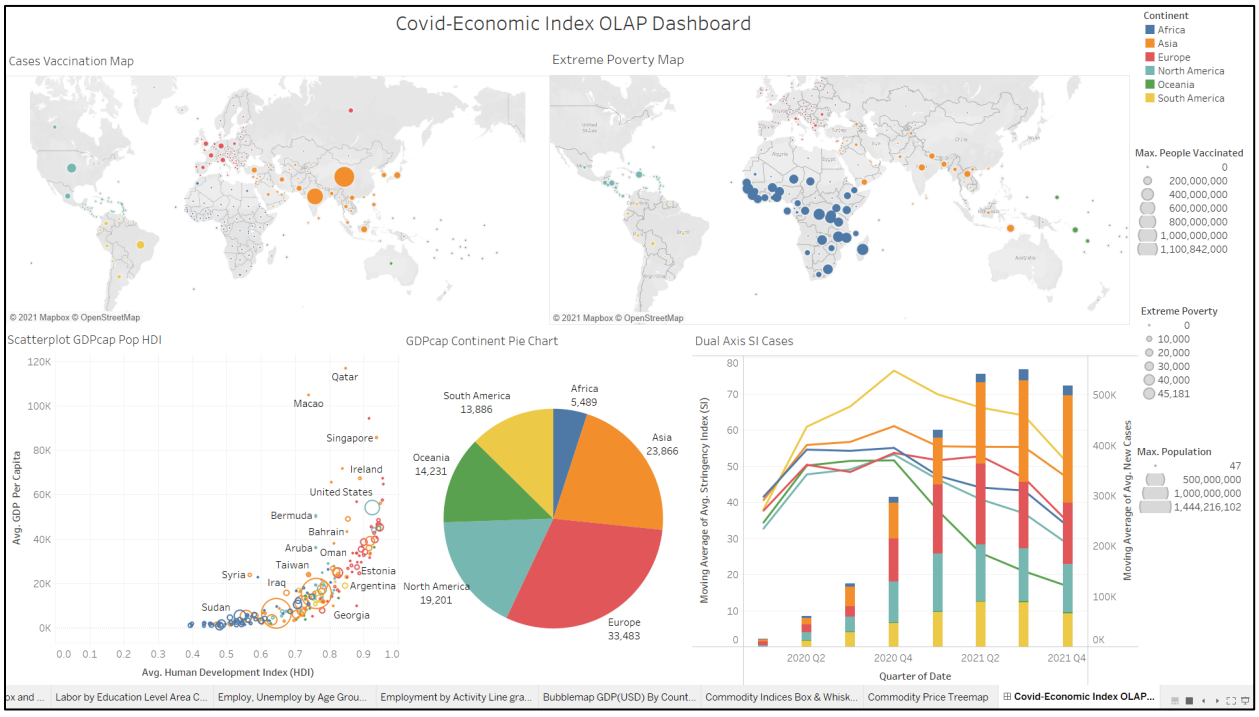


Figure 7.6: Covid-Economic Index OLAP Dashboard

According to the Covid-Economic Index OLAP Dashboard, China and India have significant people that are fully vaccinated. Meanwhile, the United States has the most Covid-19 cases. In the

symbol map, it is obvious that Africa continent has the most extreme poverty people / families recorded due to Covid-19. Most of the countries with variety number of populations has high HDI and low to moderate GDP per capita. From the visualization, there are some countries that have low population that has high HDI and high GDP per capita. In the pie chart, the Europe continent holds the biggest piece of the GDP Per capita amongst the continents. Looking into the dual combination chart, the higher the SI, the lower the cases will be. However, in the 2021 Quarter4, the cases decrease with a preferably low SI.

Analysis

From the inspection of data through a graphical view, African countries happen to have high reports on extreme poverty and low GDP per capita do not have high numbers of people that get fully vaccinated. According to research, Africa continent face low access to syringes due to the tight market and short supply (WHO Africa, 2021). The competitive syringe market might be an issue where Africa countries do not have high buying power and financial supply to import and ship the vaccinate syringes. Looking into countries with high people being vaccinated, the countries identified are India (Asia) and China (Asia). Both countries are consisted of large population by inspecting the size of circle in the scatterplot. In terms of HDI, India has low HDI while China has moderate. However, the GDP per capita of India and China is not likely to be good where India has low value and China has moderate value. In a data approach mindset, the idea came across from the observation is countries with high population is trying to boost the volume of vaccination to secure the health of the citizens and restart the economic-related sectors. It is identified countries with higher vaccination rate are continents that holds a bigger piece in the GDP Per Capita pie chart. Recommendation given for this data view is to boost the vaccination rate in the country to lift the policies for normal business operation which will contribute on the rising of GDP value.

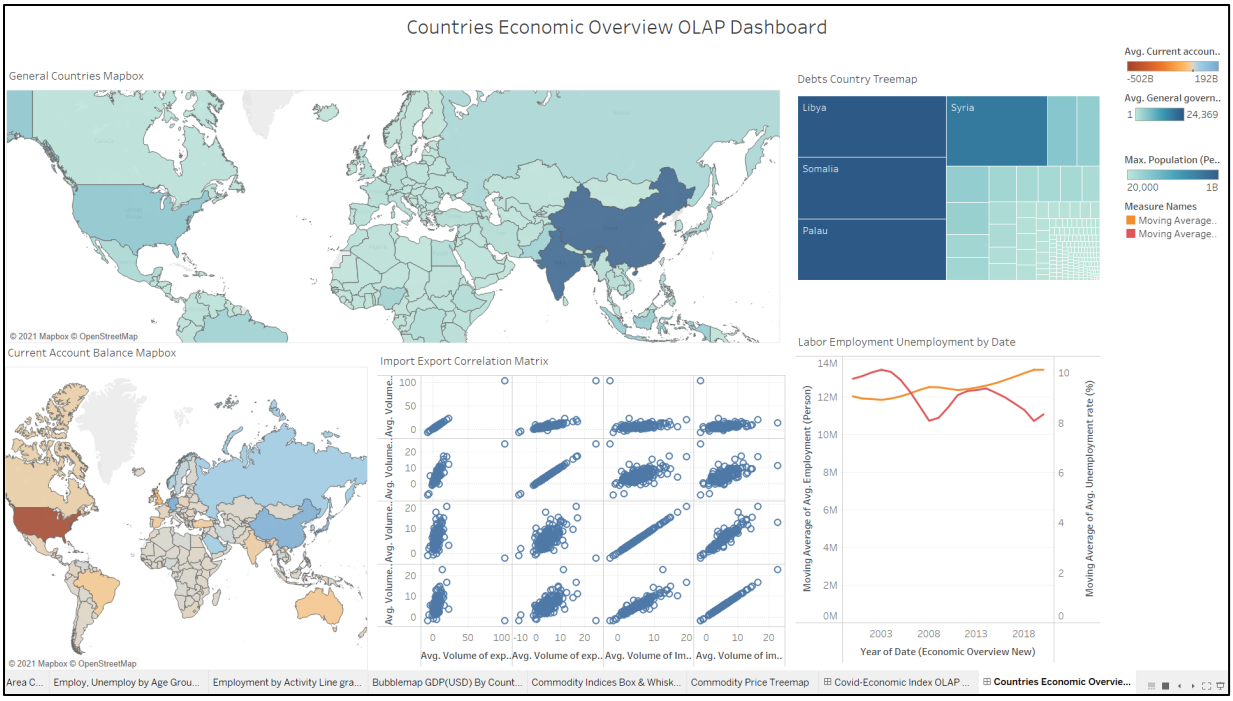


Figure 7.7: Countries Economic Overview OLAP Dashboard

The Countries Economic Overview OLAP Dashboard displays many economic indicator data in the visualization.

Analysis

From the General Countries Mapbox, countries with high population (high color density) like China and India are having high countries' revenue and expenditure. In the visualization, there is an abnormal pattern, Colombia with a less population has highest revenue and highest expenditure. To uncover the reason of the numeric figures in Colombia, Colombia is being proven to have high revenue because it is the largest South America's economic country with rich natural resources which exports gold, coffee, petroleum and coal (US News, 2020). However due to Covid-19, Colombia faced a high negative current account balance where the country practice Fiscal policy in 2020 to curb the situation. Looking into the employment and unemployment, employment value is slightly decline in 2020 while unemployment increases in 2020. The employment is not affected too throughout might be due to the rising of many new business sectors.

Note:

- Other graphs are discussed below

- Minimal OLAP

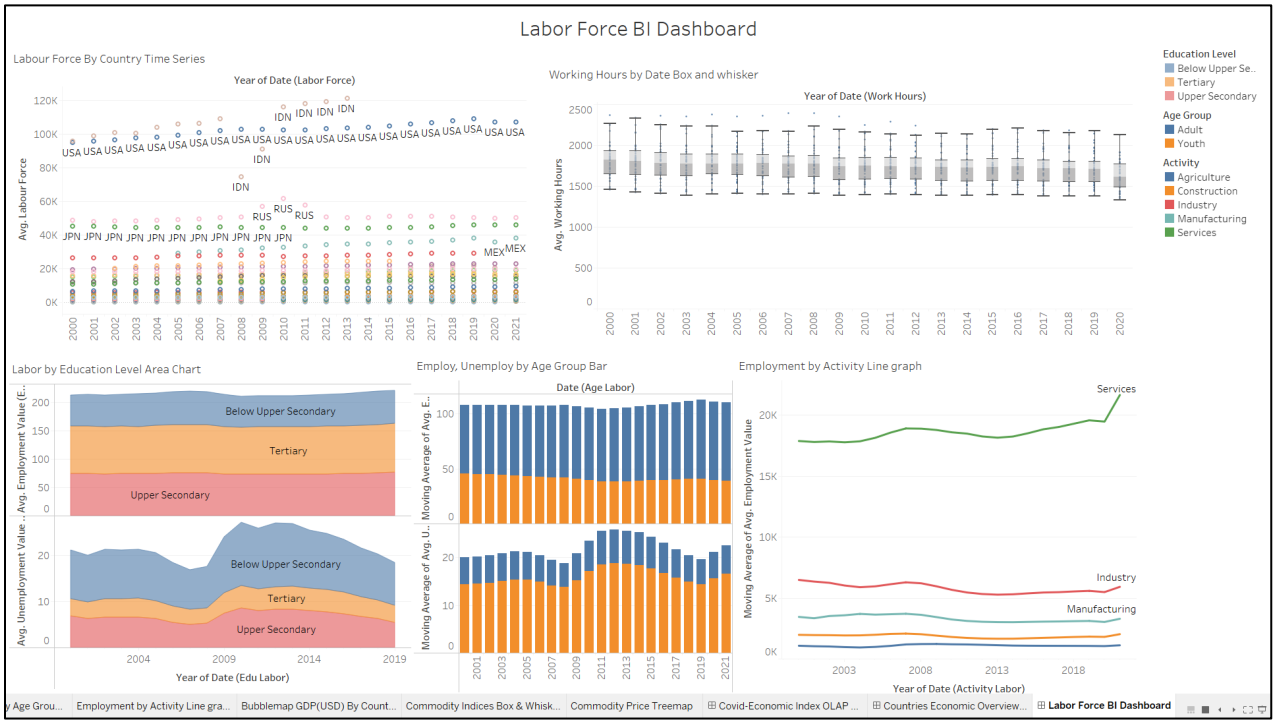


Figure 7.8: Labor Force BI Dashboard

The Labour Force BI Dashboard focuses on the number of labour force, working hours, employment and unemployment in the world.

Analysis

Looking into the box and whisker graph, the working hour during the Covid-19 period seems to have a slight drop on the value. This might be affected by the SOP practiced by countries, for instance, lockdown, reduced business hours and business closure. As a supporting evidence, Argentina (South America) implements total shutdown in the business followed by labour reduction (Orlansky & Boruchowicz, 2020). To better recover the labour force of the country. Countries are advised to encourage works that are in high demand like deliveries, IT development. At the same time, countries should focus in recovering countries resource development mainly specifying agriculture sector.

Note:

- Most graphs are being analyzed and discussed below

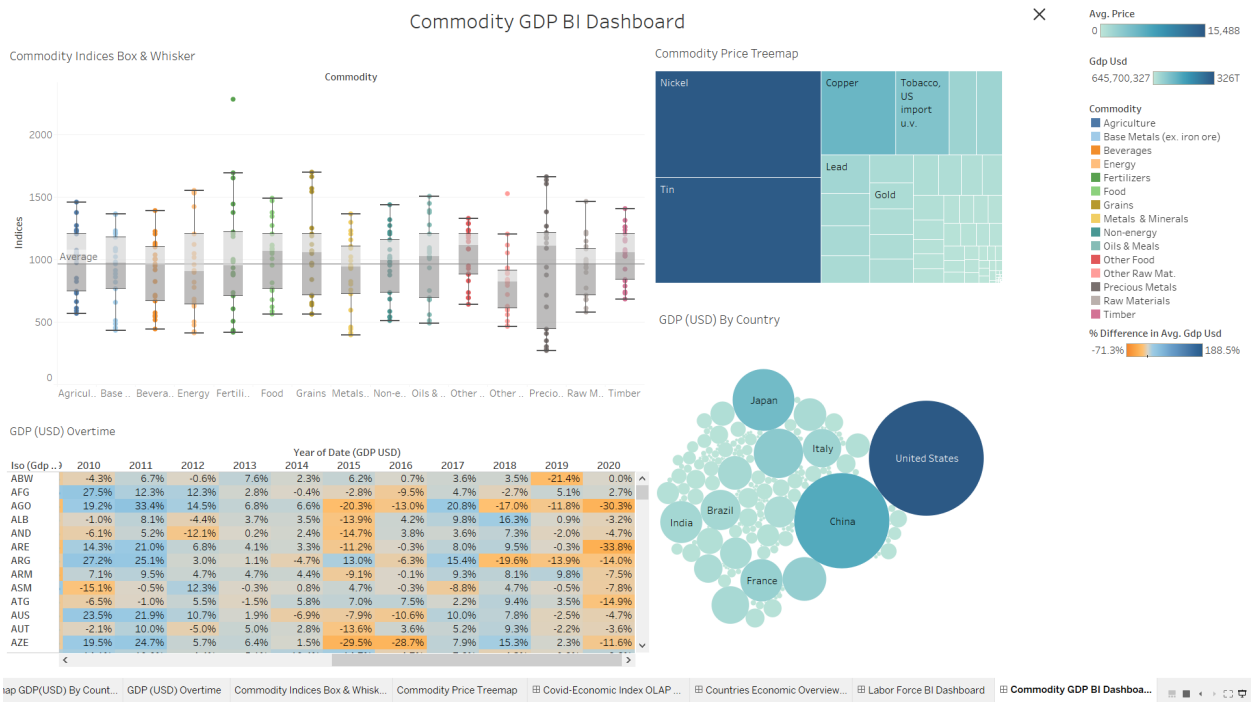


Figure 7.9: Commodity & GDP BI Dashboard

The Commodity & GDP BI Dashboard describes the commodity indices, commodity prices and the GDP (USD) owned by the countries.

Analysis

It is identified that United States has the highest GDP (USD) amongst the countries. According to the commodity price treemaps, Nickel and Tin are having a high commodity price. From this finding, a suggestion for low GDP countries is given where if the countries have high capacity on nickel and tin resources, the countries can consider exporting the resource with the high market price to maximize countries' profit. Moving on to the commodity indices, the average line of all commodity indices is 963. Based on the average line, the category of other raw material is not advisable for countries to invest on it. On the other hand, fertilizer, grain and precious material is worth to be invested by the countries due to the capabilities to reach such maximum point of the box plot. Looking into the GDP (USD) overtime by each country, most countries face decline in GDP and less countries manage to maintain or increase their GDP. This could be supporting evidence to prove that Covid-19 did impact badly on countries' economic growth.

OLAP Analysis

- **Rolling-up / Drill-down**

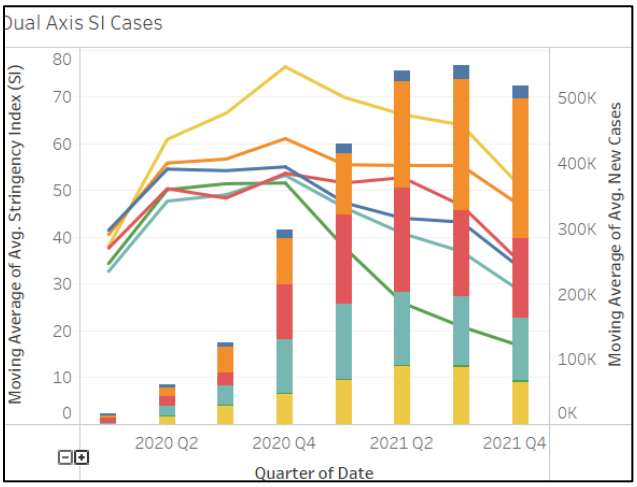
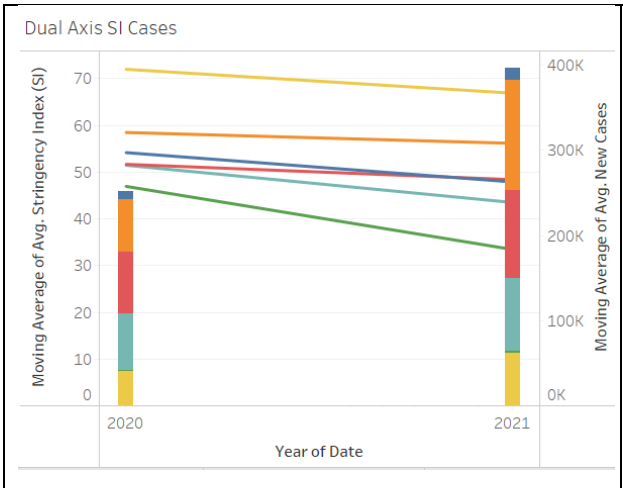
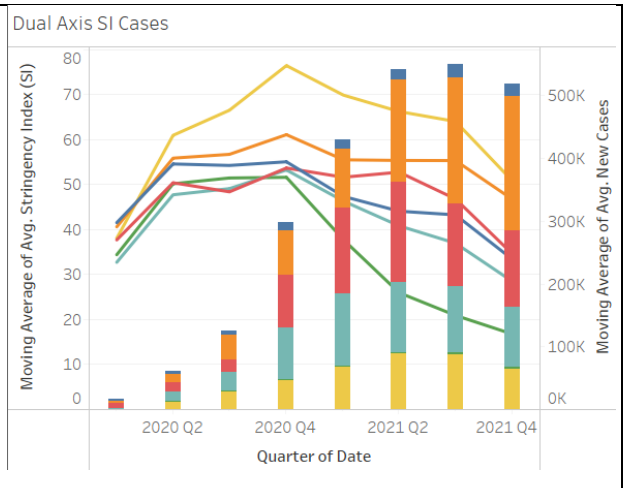


Figure 7.10: Plus / Minus Button for rolling-up & drilling-down on Date

On the OLAP dashboard, the date dimension is accessible to rolling-up and drill-down OLAP operation. The data on the visualization can be displayed by different time perspective as listed (Year > Quarter > Month >Week > Day).

Example

	
Display by Year	Display by Quarter

Analysis

It is obvious that the average of new cases (bar plot) is rising from the start of year 2020 and starting to decrease in 2021 Quarter 4. Looking on the Stringency Index (SI) changes, all continents performed strict policies noted with high SI in the start of the pandemic. According to the dual axis graph, the SI is decreasing overtime. Based on the data shown, it is being proven that the countries are getting better control on the pandemic where the lockdown policies are loosen for normal countries operation. As supporting evidence, Sydney (Australia - Oceania) had reopened the barbers, cafes and gyms on 11 October 2021 after surge of vaccinations which is equivalent to 2021 Quarter 4 (Miller, 2021). Meanwhile, in Malaysia (Asia) had announced that fully vaccinated residents are free from travel restrictions on 10 October 2021 (Chu, 2021). These findings eventually means that the countries had started to move forward to an economic recovery state since the business and travel sector is allowed through the evidence given. Recommendations toward the rolling-up/drill-down visualization is that for countries that have not effectively pull the cases down may learn from other countries' previous policies to maintain a higher SI. The South America continent is controlling the pandemic pretty good with high SI and low increase cases. According to the measures and policies imposed by the South America governments, the countries are advised to perform selective lockdown, curfew hours reduction and lifting bans on economic and education sectors for safe areas (IOM UN Migration, 2021). These actions will increase the SI and at the same time maximize economic growth for safe areas.

- **Data Slicing**

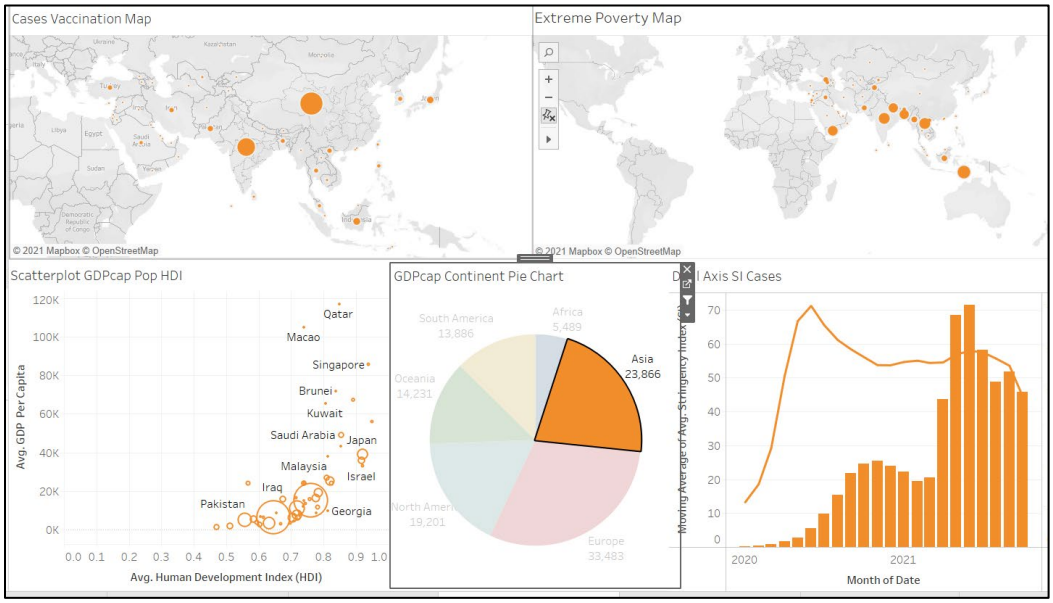


Figure 7.11: Data Slicing on Asia Continent

In the figure, data slicing is being performed on Asia Continent which is consist of several countries being visualized with several KPIs. All countries in the Asia Continent are being filtered and highlighted out.

Analysis

The Asia Continent holds the second largest average of GDP Per Capita amongst the continents. From the scatterplot, most Asian countries have moderate HDI and moderate / low GPD. Still, in the scatterplot, Asian countries with low population such as Qatar and Macao have a high HDI and high GDP Per Capita. A brief idea on relationship of population and GDP can be made. It is suggested that high population will not have high GDP per capita as seen in the data both China and India with high population are not the peak countries that have high GDP. During the time where Asia is having high SI, the cases are being well controlled. However, when the SI decrease, the new cases are increased which explains the important of practicing standard operation policies during Covid-19 to control the spread of virus.

- **Data Dicing**

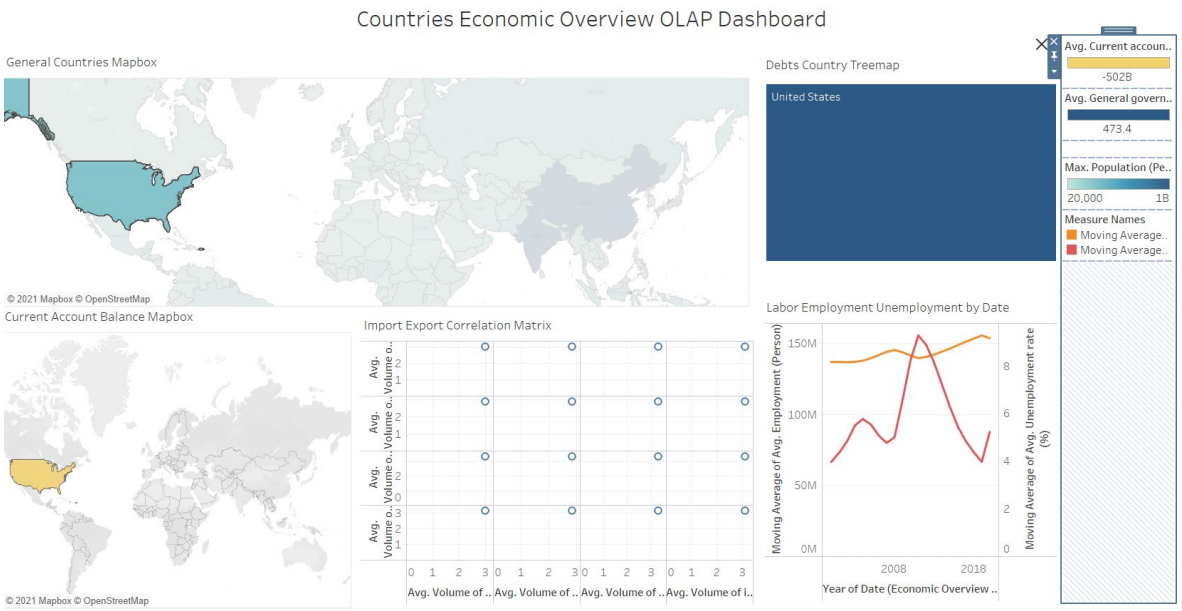


Figure 7.12: Data Dicing on United States (North America Continent)

Data dicing is being performed on United States country where all relevant metrics are displayed on only the US country. This dices the geographical data into the North America continent which furthering into United States only.

Analysis

US as a strong financial power country that has preferably more than average population is facing a high amount of negative current account balance. By referring the latest news, the US is facing bad economic status where the country is encountering surging product prices, labor shortage and messy supply chain which explains the negative figure of current account balance in the country impacted by Covid-19 pandemic (Goldman & Tappe, 2021). However, the US do not have high government debt by referring the value on the treemaps. From the matrix, US is identified to have high imports and exports transaction on goods and services. Through research, the US economy face a great crisis due to the pandemic where demand shock, supply shock and monetary shock all happen at once. The social distancing, lockdown and isolation in US had precipitated severe economic downturn as economy's capacity in goods and services production is being reduced badly (Bauer, Broady, Edelberg, & O'Donnel, 2020)

Data Distributions (Graph)

- Maps

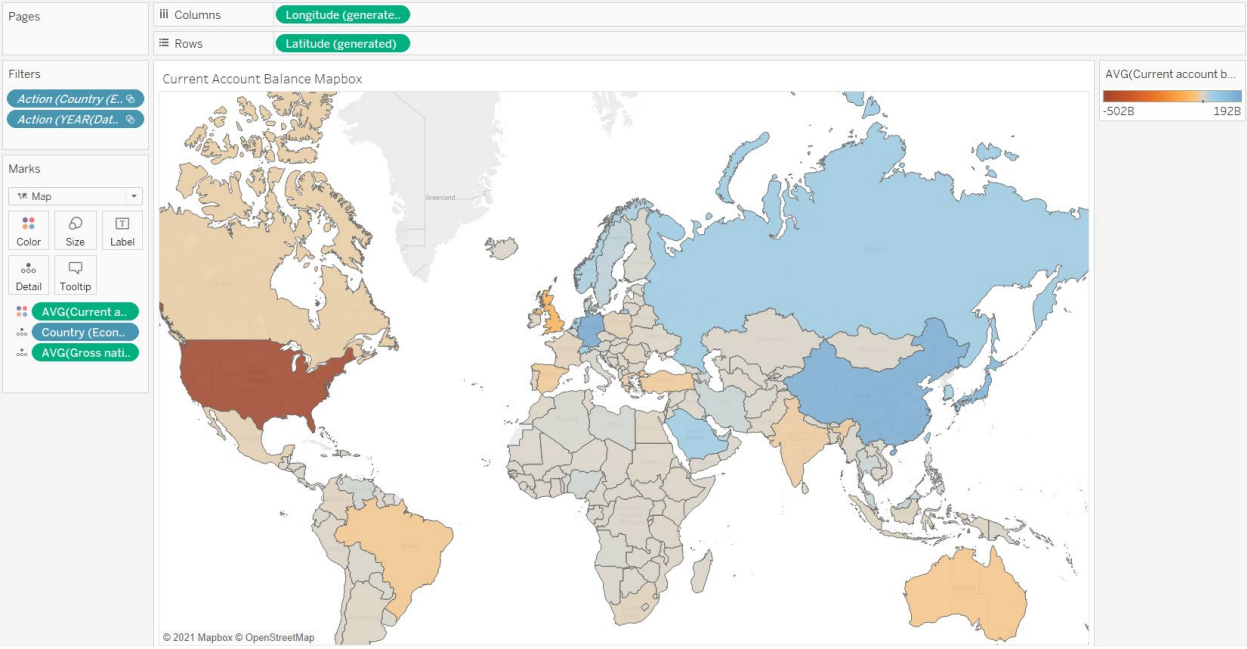


Figure 7.13: Average Current Account Balance, Gross nation savings by Countries Map

The maps graphical presentation is being selected to display the countries on the map based on the system generated longitude and latitude. The country / region data is being displayed geographically with the colour identification on the amount of average current account balance in the countries. In the details while hovering, researcher can obtain details on countries name, current account balance, gross nation savings. To draw a small analysis on the graph, from the colour gradients, countries that have high positive average current account balance are China, Japan, Russia, Germany and Saudi Arabia. By observing the map, the countries with high average current account balance do not define as countries that have high gross nation savings. From all the listed countries, only China has a high gross nation savings.

- Symbol Maps

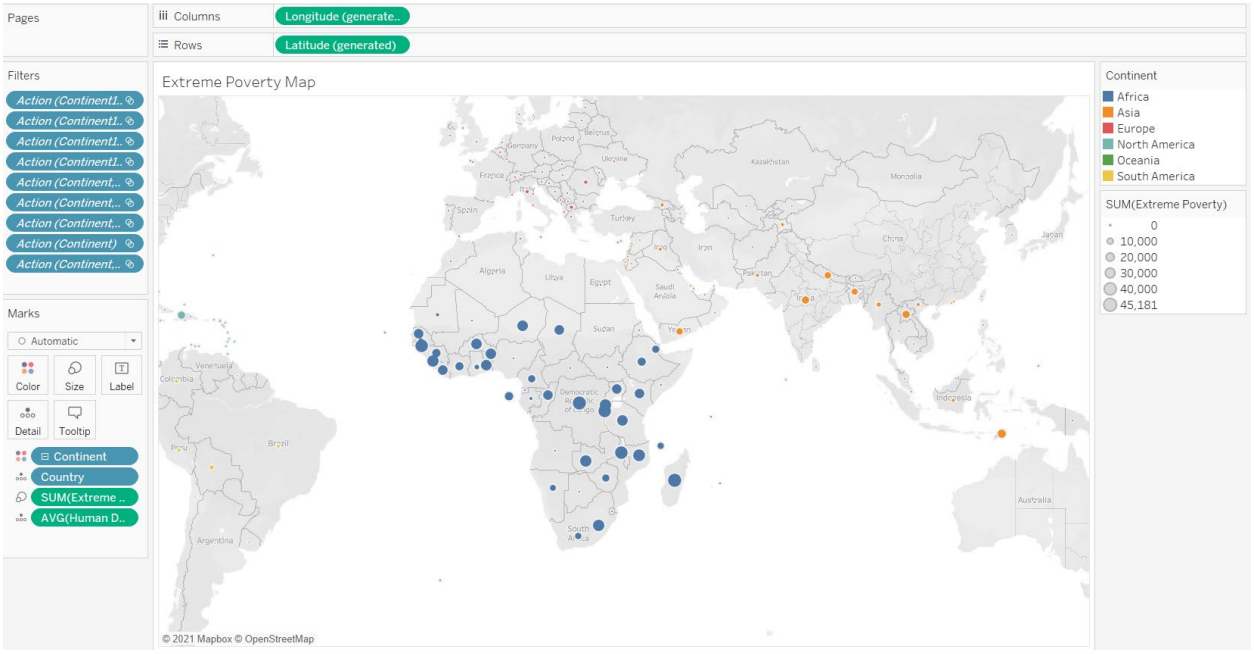


Figure 7.14: Extreme Poverty Symbol Maps

The Extreme Poverty Symbol Maps display the amount of extreme poverty reports grouped by the country. The countries are being colored with the same continent and having the detail of average human development. The size of the circle explains the level of extreme poverty reports total in the country throughout the covid period. Based on the data distribution on the symbol maps, most countries in Africa reported high volume of extreme poverty cases throughout the Covid-19 period. In Asia too, there are some appealing circles which indicates the cases of extreme poverty. By inspecting the data, the countries that happen to have quite an amount of extreme poverty cases are having either moderate or low (most of the identified countries) HDI. The low standard in living is somehow relates to the reason behind of the reported poverty cases.

- Pie Charts

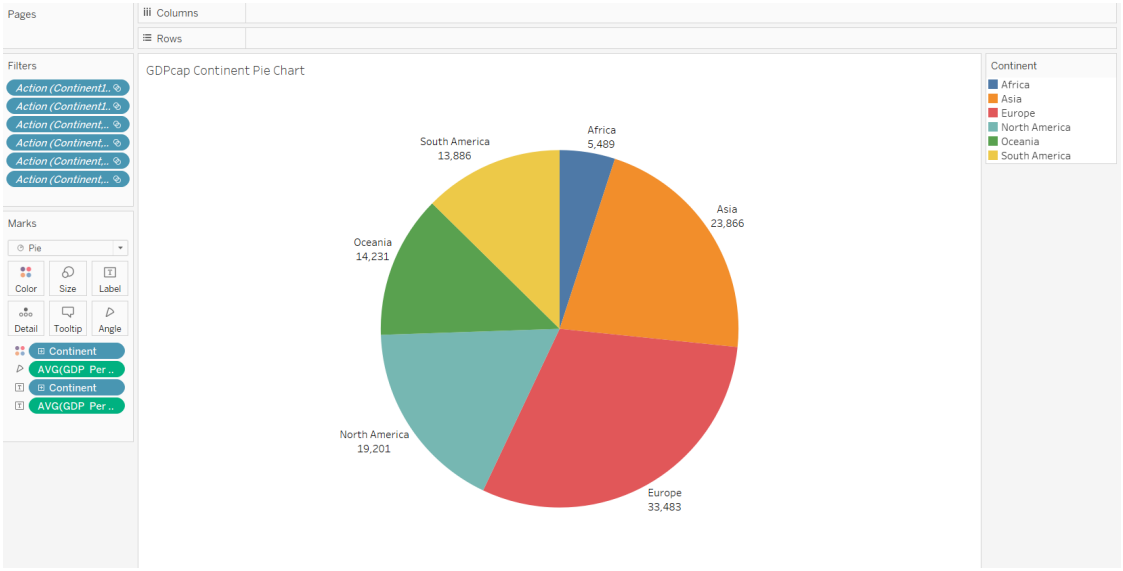


Figure 7.15: GDP Per Capita by Continents Pie Chart

The pie chart above displays the GDP Per Capita segmented by continents where the colour is applied accordingly. Europe is the continent that has the highest GDP Per Capita which is then followed by Asia, North America, South America, Oceania and Africa. As a recommendation, to improve the countries' GDP, countries in South America and Oceania are advised to learn from the countries in Europe for better financial allocation and quality of goods and services.

- Stacked Bars

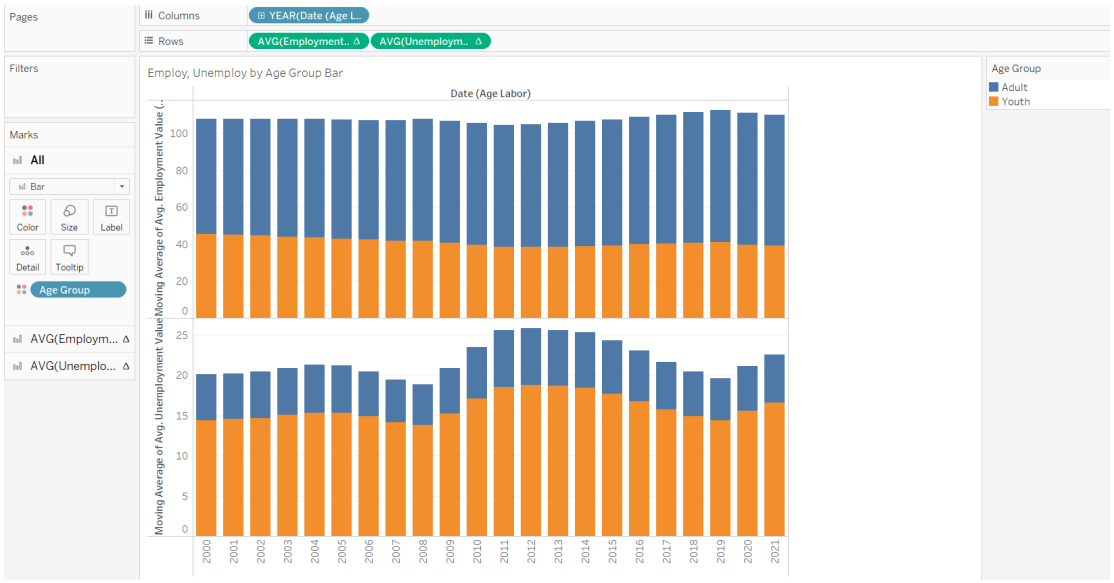


Figure 7.16: Average of employment and unemployment by Age Group Overtime

The stacked bars graph shows the average of employment value and average of unemployment value being grouped by age group. In terms of the employment value, both youth and adult age groups do not have many changes Focusing on the timestamp from 2019 to 2021 (Covid-19 pandemic), there is a slight decline for employment value for adult and youth. On the other hand, the unemployment is increasing for both age groupings at the Covid-19 period. By referring the news, Malaysia (Asia) recorded a raise on unemployment rate and decrease on employment rate to curb with the Covid-19 pandemic on 9th August 2021 (The Star, 2021).

- Treemaps

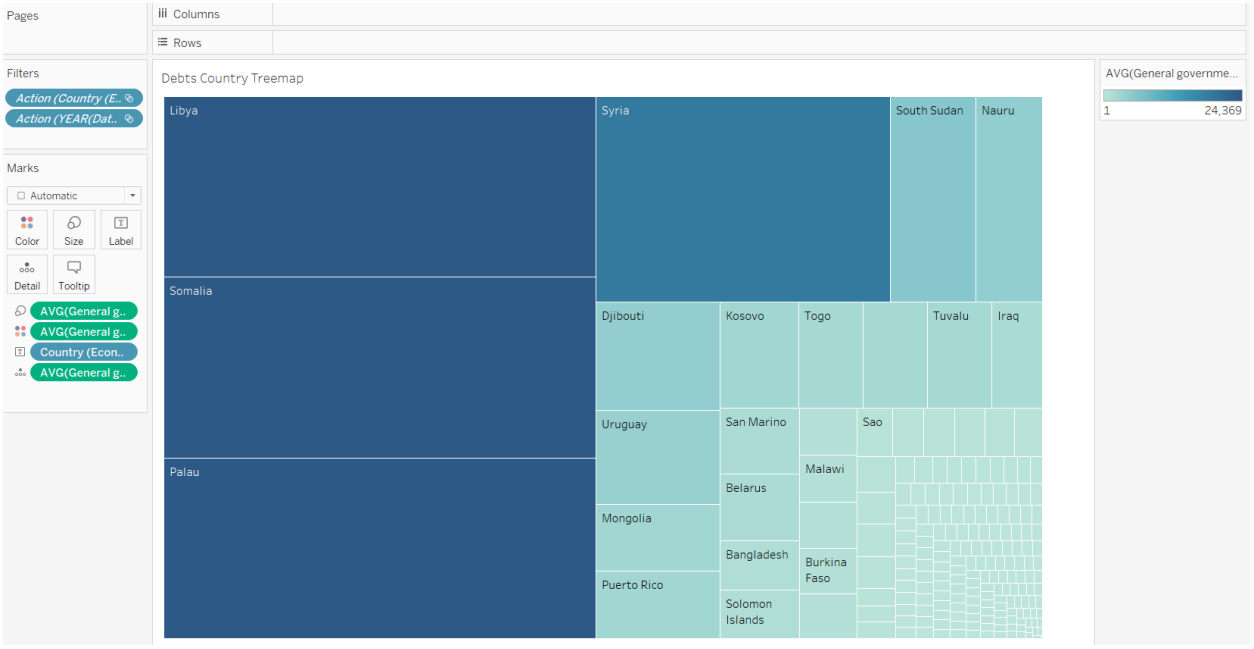


Figure 7.17:Debts, Inflation by Countries Treemaps

The treemaps on countries' general government debt is created where the density of the block color indicates the degree of the average debt value. Through visualization, countries such as Palau, Libya and Somalia have high government debt. These countries are advised to work on alliance with other country and try to produce quality products. During the period of Covid-19, Somalia's request on $5billion dollars assistance is approved by the International Monetary Fund and World Bank The slash of the debts help much in the Somalia to curb with the Covid-19 pandemic with their hard work on reforming and recover the development of the country (Ahmed, 2020).

- Circle Views

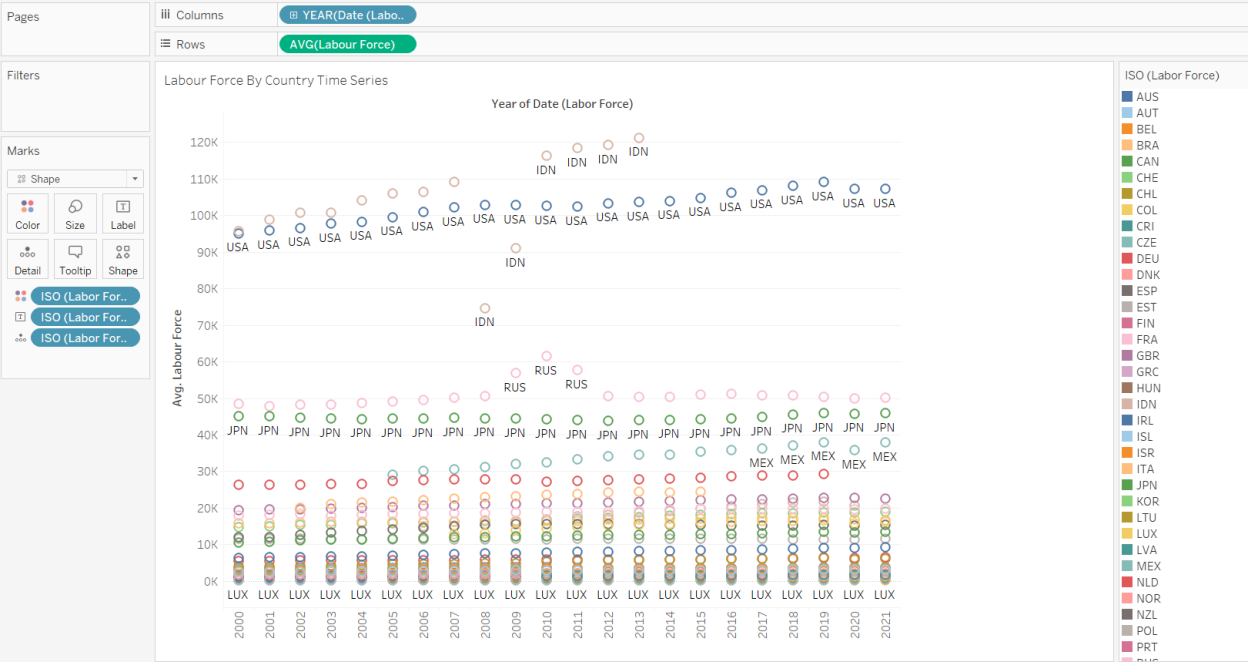


Figure 7.18: Average of labor force by country circle views

The circle views graph is implemented on the distribution on country ISO code in inspecting the average of labour force in a yearly basis. During the normal years without Covid-19, the labour force of most countries is either steadily increase or slightly increasing except for Indonesia facing a decline in 2008 and 2009). Specifying the time frame during Covid-19, in 2020, Mexico and USA experienced a decline in labour force which indirectly explains that there are unemployment happening due to Covid-19. To strengthen the idea obtained from the data visualization, Mexico is proven to have a decrease in labour force where the countries had shed 12.5 million occupations in the beginning of Covid-19 crisis (UBS, 2021)

- Lines (Continuous)

Figure 7.19: Lines (Continuous) on Employment by Activity Yearly basis

The continuous lines graph is utilized to present the average employment value by activity in a yearly basis. Services as the commonly provided business activity to consumer has a high employment value. This data view is supported by the demand in food delivery service during Covid-19. To support the statement, restaurants and dining business places are being forced to close during the lockdowns, ride-hailing business company like Uber generate more revenue than usual (Keane, 2020). There are a slight increase on Industry, Manufacturing and Construction sectors. The Industry and manufacturing industry increase because there is a high demand in face mask and hand sanitizer. To backup the idea, luxury brand company, LVMH switched their production line to producing hand sanitizer and hygienic masks in 2020 (Betti & Heinzmann, 2020). Meanwhile, construction is affected as well because there is a need to build quarantine center to ease the control on virus outbreak. For example, China had constructed 5000 room quarantine facility in Guangzhou for overseas arrivals (Gan & Yeung, 2021).

- Area Charts (Continuous)

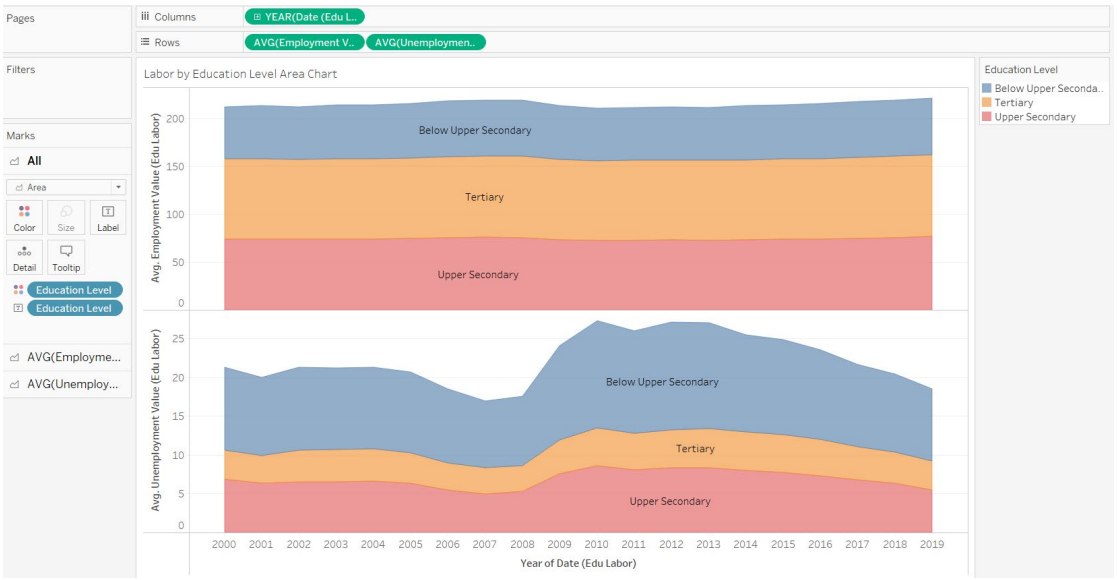


Figure 7.20: Employ, Unemploy by Education level Yearly basis

The area charts display the employment and unemployment for different education level. However, the changes on employment value is not much likely change before the Covid-19 pandemic as 2019 is the beginning of Covid-19 to be discovered. In the unemployment value growth, there are not many insights obtained to assist in determining the impact of Covid-19 to labour force. The conclusion can be obtained from this graph is in the beginning phase of Covid-19, the labour force is not affected.

- Dual Combination

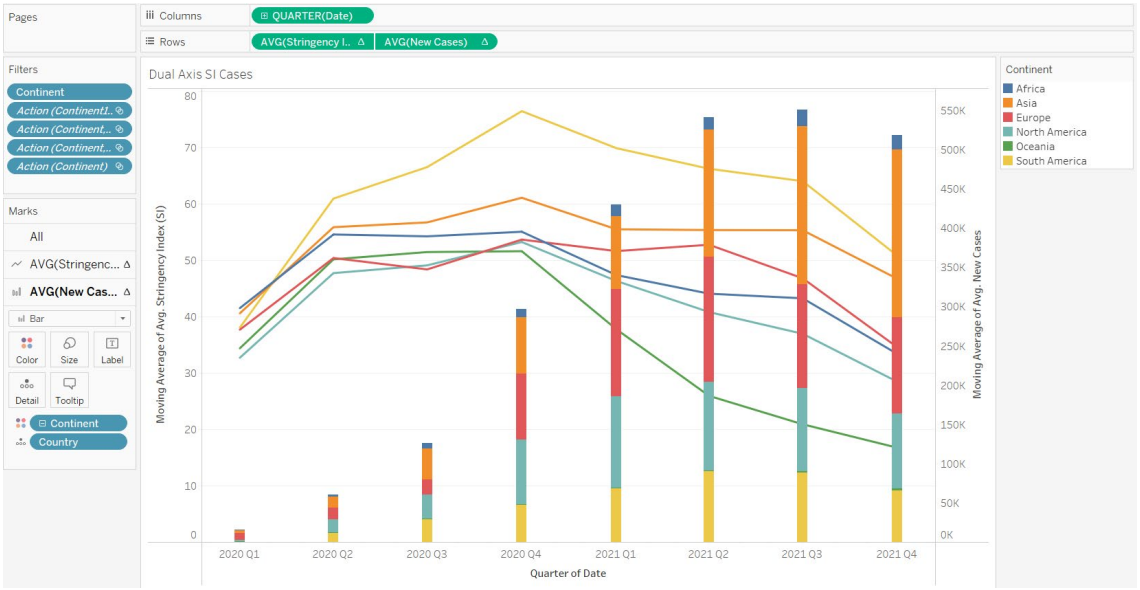


Figure 7.21: SI, New Cases Dual Combination by Continent quarterly basis

The dual combination of line and bar axis is being used to plot the SI and new cases by continent controlled by date.

Explained on previous pages.

- Scatter Plots

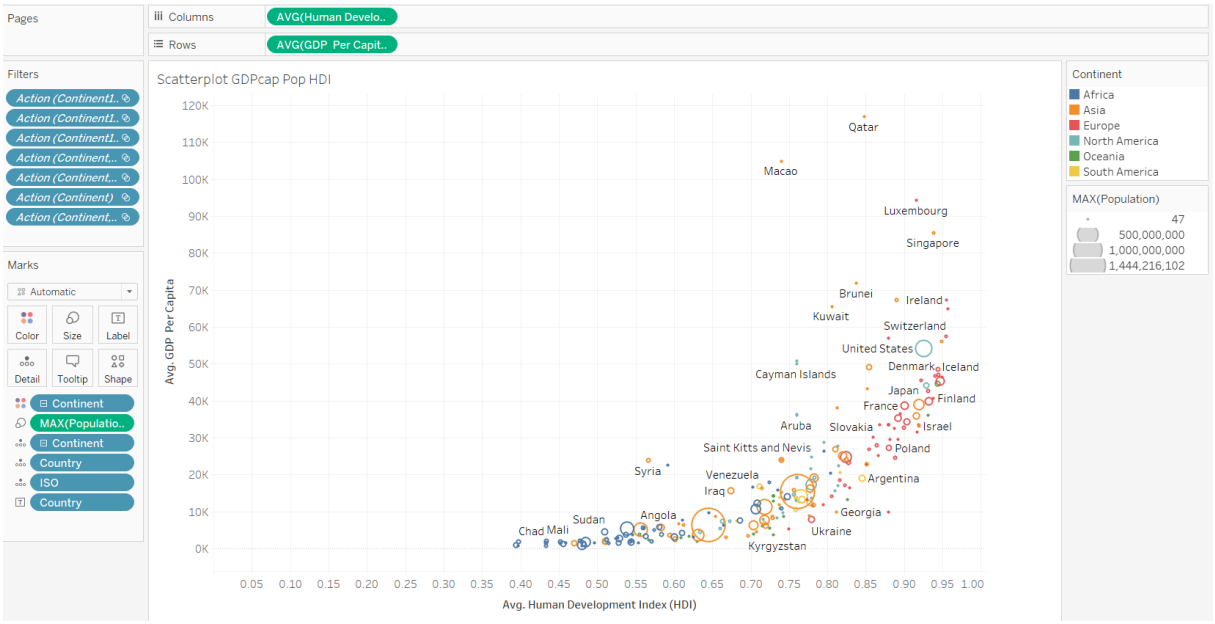


Figure 7.22: Scatterplot on GDP per capita, population, HDI by Countries

Scatterplot displays the population, GDP Per Capita and HDI plotted by the countries.

Explained on previous page

- Box-and-whisker Plots

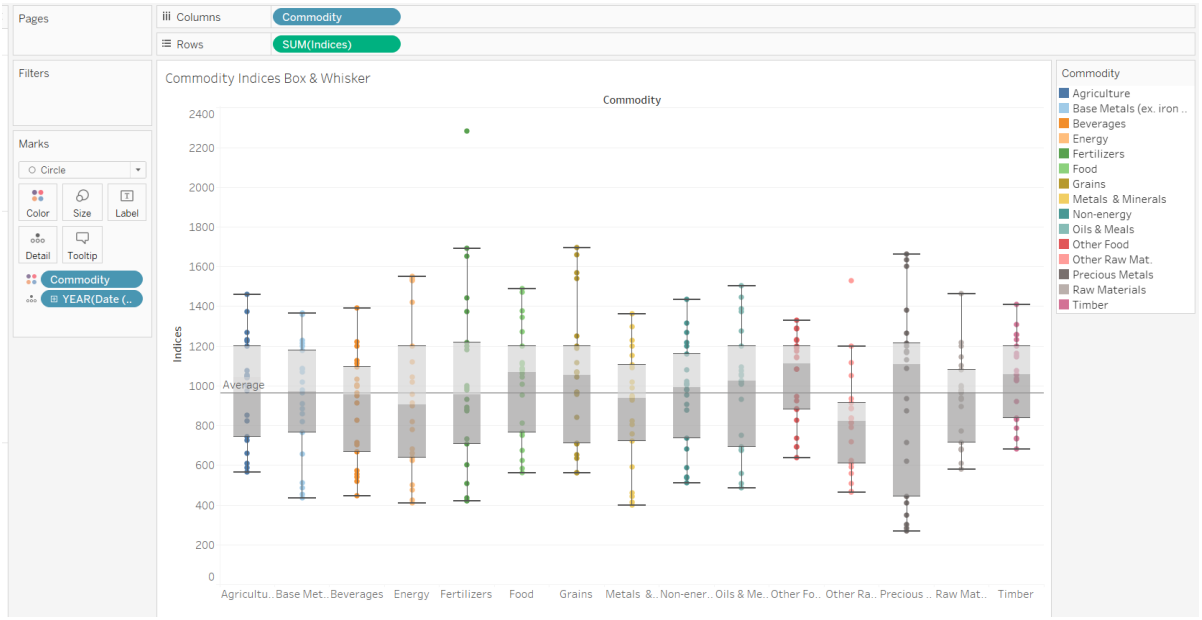


Figure 7.23: Commodity Indices Box & Whisker Plot

Box and whisker plot visualizes the commodity indices based on the type of commodity type. Each box and whiskers are coloured with respective commodity type.

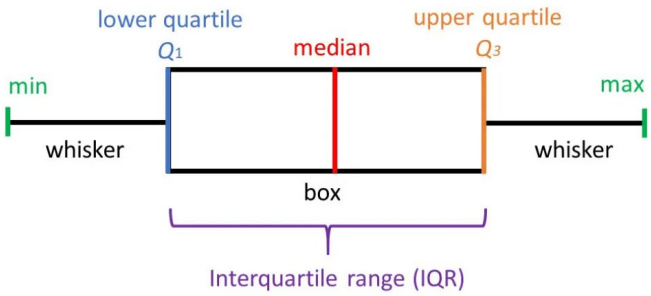


Figure 7.24: Box and Whisker Structure
(McLeod, 2019)

Through the understanding box and whisker structure, the box plots in the statistic seem to have similar upper quartile, Q3. The precious material has the longest interquartile range (IQR) where the median of the box is way beyond the average line of the graph.

- Packed Bubbles

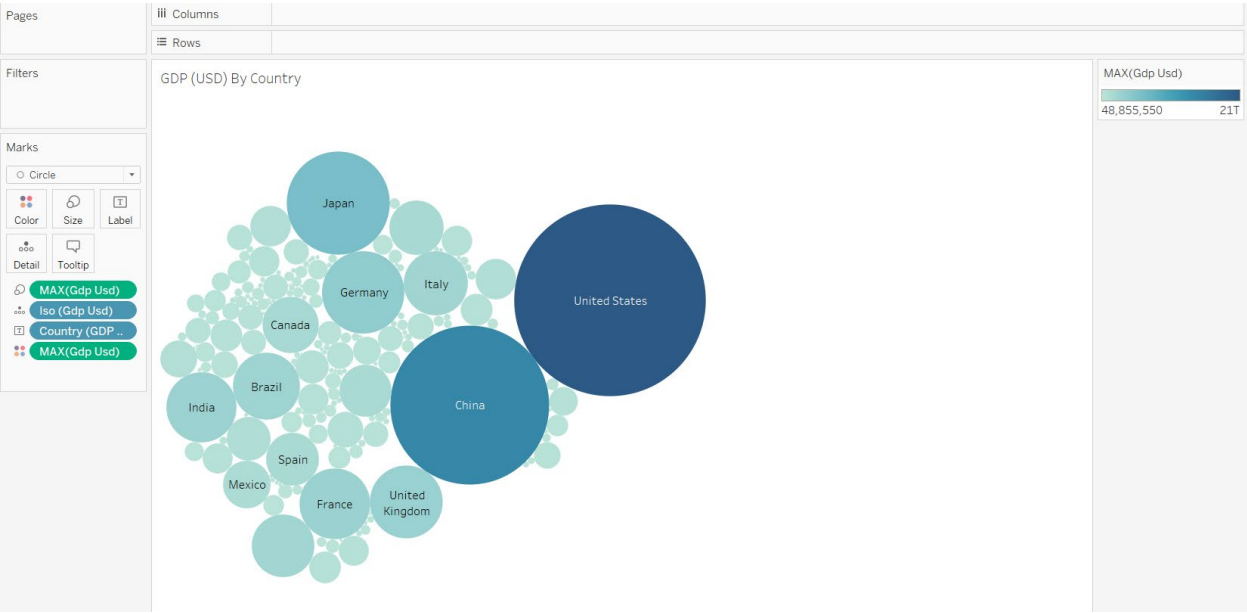
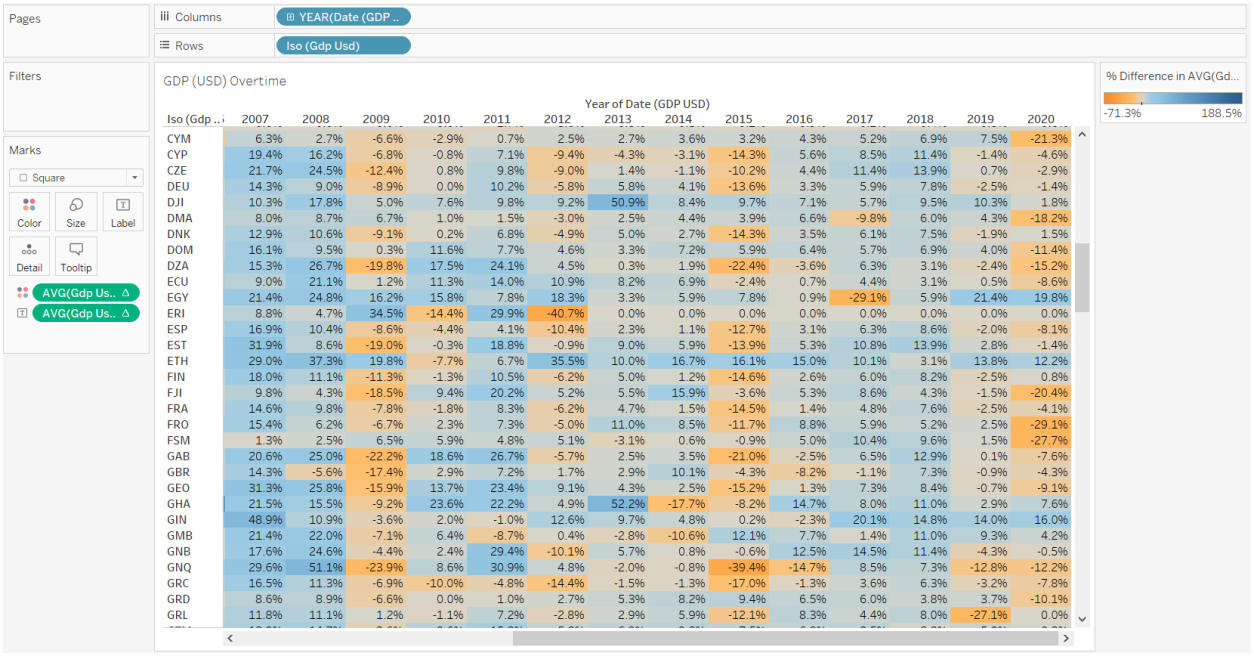


Figure 7.25:GDP (USD) by Country Packed Bubbles

Packed bubbles graph is used to visualize the data distribution of GDP (USD) by Country. The bigger the size of bubbles represents, the higher the amount of the maximum of GDP (USD) of the countries. Obviously, United States has the highest GDP (USD) which is followed by China and Japan. The visualization showcases the economic power of each country where details of ISO code, country name, maximum of GDP calculated in USD ($) is being shown. The high GDP of country eventually indicates high progress of countries development measured by currency value.

- Highlight Tables



Iso (Gdp ...	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
CYM	6.3%	2.7%	-6.6%	-2.9%	0.7%	2.5%	2.7%	3.6%	3.2%	4.3%	5.2%	6.9%	7.5%	-21.3%
CYP	19.4%	16.2%	-6.8%	-0.8%	7.1%	-9.4%	-4.3%	-3.1%	-14.3%	5.6%	8.5%	11.4%	-1.4%	-4.6%
CZE	21.7%	24.5%	-12.4%	0.8%	9.8%	-9.0%	1.4%	-1.1%	-10.2%	4.4%	11.4%	13.9%	0.7%	-2.9%
DEU	14.3%	9.0%	-8.9%	0.0%	10.2%	-5.8%	5.8%	4.1%	-13.6%	3.3%	5.9%	7.8%	-2.5%	-1.4%
DJI	10.3%	17.8%	5.0%	7.6%	9.8%	9.2%	50.9%	8.4%	9.7%	7.1%	5.7%	9.5%	10.3%	1.8%
DMA	8.0%	8.7%	6.7%	1.0%	1.5%	-3.0%	2.5%	4.4%	3.9%	6.6%	-9.8%	6.0%	4.3%	-18.2%
DNK	12.9%	10.6%	-9.1%	0.2%	6.8%	-4.9%	5.0%	2.7%	-14.3%	3.5%	6.1%	7.5%	-1.9%	1.5%
DOM	16.1%	9.5%	0.3%	11.6%	7.7%	4.6%	3.3%	7.2%	5.9%	6.4%	5.7%	6.9%	4.0%	-11.4%
DZA	15.3%	26.7%	-19.8%	17.5%	24.1%	4.5%	0.3%	1.9%	-22.4%	-3.6%	6.3%	3.1%	-2.4%	-15.2%
ECU	9.0%	21.1%	1.2%	11.3%	14.0%	10.9%	8.2%	6.9%	-2.4%	0.7%	4.4%	3.1%	0.5%	-8.6%
EGY	21.4%	24.8%	16.2%	15.8%	7.8%	18.3%	3.3%	5.9%	7.8%	0.9%	-29.1%	5.9%	21.4%	19.8%
ERI	8.8%	4.7%	34.5%	-14.4%	29.9%	-40.7%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
ESP	16.9%	10.4%	-8.6%	-4.4%	4.1%	-10.4%	2.3%	1.1%	-12.7%	3.1%	6.3%	8.6%	-2.0%	-8.1%
EST	31.9%	8.6%	-19.0%	-0.3%	18.8%	-0.9%	9.0%	5.9%	-13.9%	5.3%	10.8%	13.9%	2.8%	-1.4%
ETH	29.0%	37.3%	19.8%	-7.7%	6.7%	35.5%	10.0%	16.7%	16.1%	15.0%	10.1%	3.1%	13.8%	12.2%
FIN	18.0%	11.1%	-11.3%	-1.3%	10.5%	-6.2%	5.0%	1.2%	-14.6%	2.6%	6.0%	8.2%	-2.5%	0.8%
FJI	9.8%	4.3%	-18.5%	9.4%	20.2%	5.2%	5.5%	15.9%	-3.6%	5.3%	8.6%	4.3%	-1.5%	-20.4%
FRA	14.6%	9.8%	-7.8%	-1.8%	8.3%	-6.2%	4.7%	1.5%	-14.5%	1.4%	4.8%	7.6%	-2.5%	-4.1%
FRO	15.4%	6.2%	-6.7%	2.3%	7.3%	-5.0%	11.0%	8.5%	-11.7%	8.8%	5.9%	5.2%	2.5%	-29.1%
FSM	1.3%	2.5%	6.5%	5.9%	4.8%	5.1%	-3.1%	0.6%	-0.9%	5.0%	10.4%	9.6%	1.5%	-27.7%
GAB	20.6%	25.0%	-22.2%	18.6%	26.7%	-5.7%	2.5%	3.5%	-21.0%	-2.5%	6.5%	12.9%	0.1%	-7.6%
GBR	14.3%	-5.6%	-17.4%	2.9%	7.2%	1.7%	2.9%	10.1%	-4.3%	-8.2%	-1.1%	7.3%	-0.9%	-4.3%
GEO	31.3%	25.8%	-15.9%	13.7%	23.4%	9.1%	4.3%	2.5%	-15.2%	1.3%	7.3%	8.4%	-0.7%	-9.1%
GHA	21.5%	15.5%	-9.2%	23.6%	22.2%	4.9%	52.2%	-17.7%	-8.2%	14.7%	8.0%	11.0%	2.9%	7.6%
GIN	48.9%	10.9%	-3.6%	2.0%	-1.0%	12.6%	9.7%	4.8%	0.2%	-2.3%	20.1%	14.8%	14.0%	16.0%
GMB	21.4%	22.0%	-7.1%	6.4%	-8.7%	0.4%	-2.8%	-10.6%	12.1%	7.7%	1.4%	11.0%	9.3%	4.2%
GNB	17.6%	24.6%	-4.4%	2.4%	29.4%	-10.1%	5.7%	0.8%	-0.6%	12.5%	14.5%	11.4%	-4.3%	-0.5%
GNQ	29.6%	51.1%	-23.9%	8.6%	30.9%	4.8%	-2.0%	-0.8%	-39.4%	-14.7%	8.5%	7.3%	-12.8%	-12.2%
GRC	16.5%	11.3%	-6.9%	-10.0%	-4.8%	-14.4%	-1.5%	-1.3%	-17.0%	-1.3%	3.6%	6.3%	-3.2%	-7.8%
GRD	8.6%	8.9%	-6.6%	0.0%	1.0%	2.7%	5.3%	8.2%	9.4%	6.5%	6.0%	3.8%	3.7%	-10.1%
GRL	11.8%	11.1%	1.2%	-1.1%	7.2%	-2.8%	2.9%	5.9%	-12.1%	8.3%	4.4%	8.0%	-27.1%	0.0%

Figure 7.26: Average GDP (USD) by Country Yearly Highlight Table

The average GDP changes by countries is being presented through highlight table which showcase the data in a yearly basis. Based on the percent differences on GDP (USD) of all countries, most countries face huge decline in the countries' GDP. For example, in UK (ISO: GBR), the country faces a -4.3% of GDP (USD) in 2020. According to a research briefing, the decline in GDP of UK is the first lockdown and further lockdowns during the autumn and winter (Harari, Keep, & Brien, 2021). To curb with the high inflation in GDP, the countries can decide in implementing contractionary monetary policy which reduce bond loan prices and raise the countries' interest rates. This policy assists the countries in reduction spending where limited access in funds will endorse people to save money and low expenditure eventually halt economic growth (Kramer, 2021).

Outlier Detections

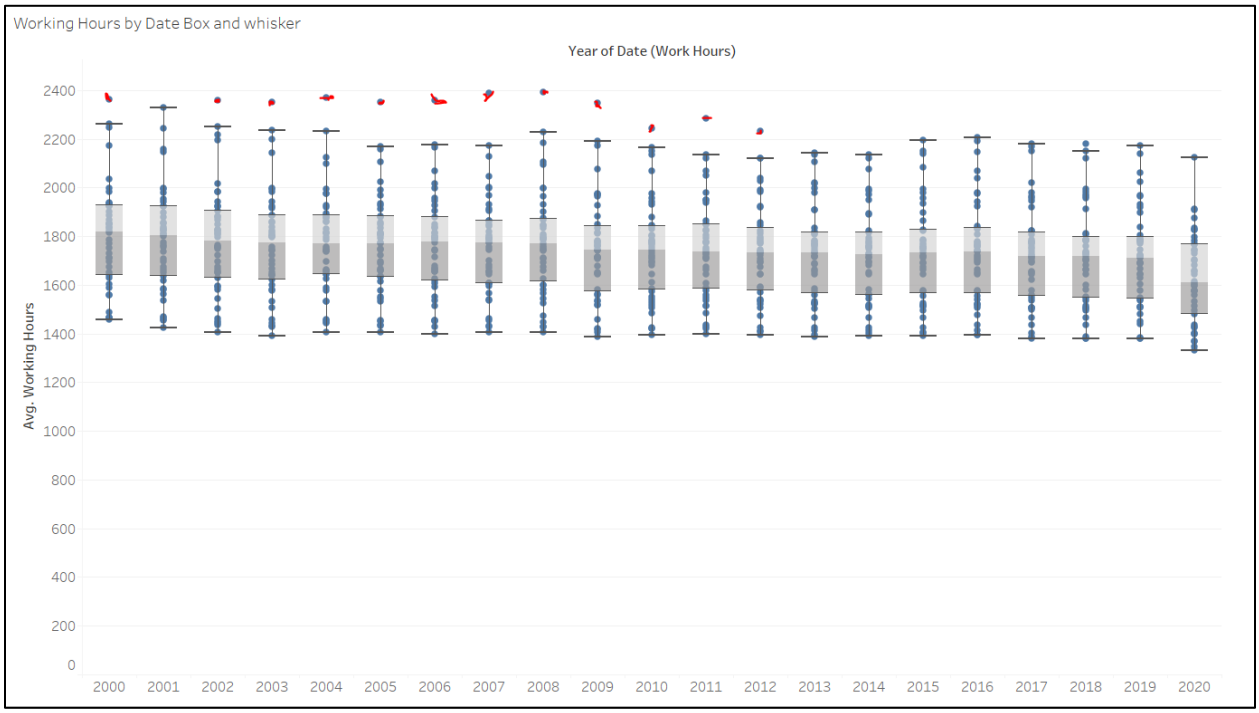


Figure 7.27: Outliers (Red) in Box and Whisker Graph

In the box and whisker plots, it is obvious that there are outliers in the data which are presented as a plot that falls outside of the box. The definition of outlier is the abnormal value that place quite a distance from other data values in the sample data. The graph is identified to be having mild outliers (highlighted in red) which is located either beyond or below the inner fence of the box (Nist/Sematech, 2012). After the investigation on the data values that are identified as outliers, there is no need of filtering out the values because they are meaningful information. The reason why they are being placed out of the fences is that the countries (e.g. Costa Rica) have relatively higher average working hours compare to the rest of the countries. Therefore, the outliers are being remained on the graph for an accurate data evaluation without removing any important data values.

Correlation Matrix

Figure 7.28: Correlation Matrix on Countries Import Export

The correlation matrix / scatter matrix on countries import export is being visualized to understand the relationship between import and export changes in the countries. A scatter matrix is an estimation of data covariance with the same number of rows and columns matrix which provides the information about how two data variables interact in the sense of direction and magnitude (Raghavan, 2018). The computation of the covariance is through the formula of:

$$S = \sum_{k=1}^{n} (\boldsymbol{x}_k - \boldsymbol{m})\,(\boldsymbol{x}_k - \boldsymbol{m})^T$$

where $\boldsymbol{m}$ is the mean vector

$$\boldsymbol{m} = \frac{1}{n} \sum_{k=1}^{n} \boldsymbol{x}_k$$

In analysis, the matrix box where same variable is being compared will be ignored. According to the matrix, the export changes have a steady increasing relationship which is logically acceptable. However, the relationship between both imports does not interpret clear relationship as well as imports and exports. In short, the pattern of correlated import & import, export & import do not have direct relationship.

Task 2: Clustering

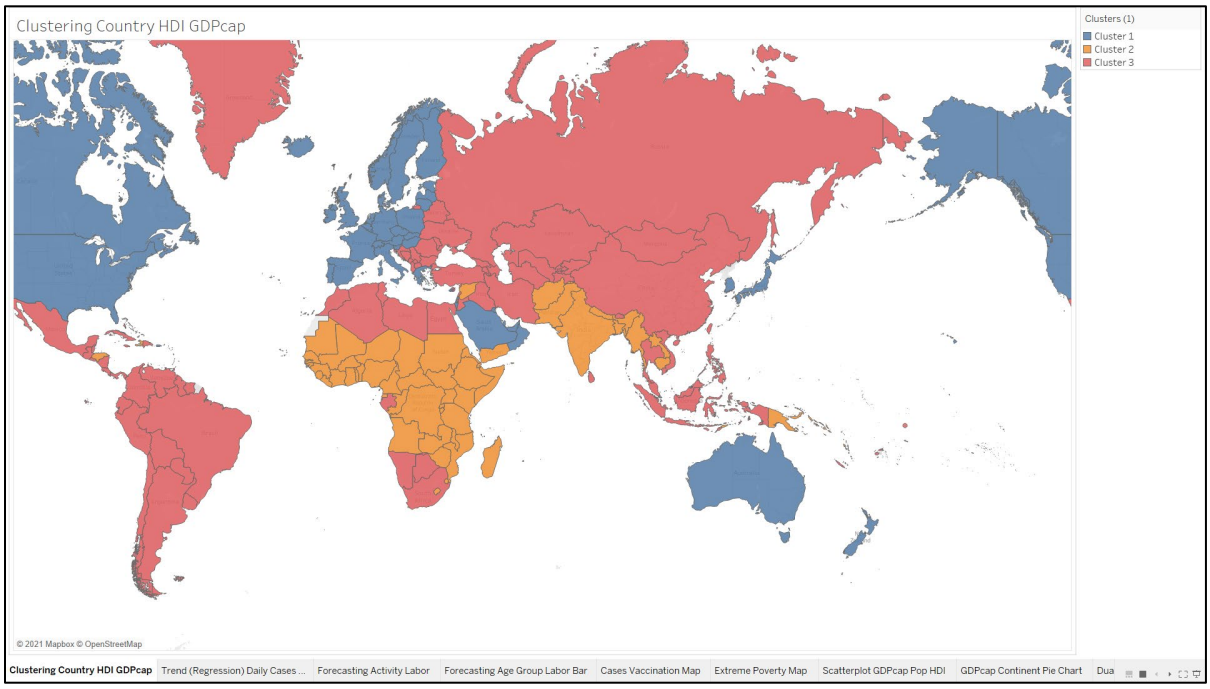


Figure 7.29: Clustering on Human Development Index (HDI), GDP per Capita by Country

Clustering is an unsupervised learning method that segments data based on similarity on same data variable values. The clustering algorithm applied for the data research is k-means where data are partitioned into number of k clusters. The k-means algorithm determines the grouping of data by initializing centroids (mean value of all data points in the cluster) iteratively and appoint the data to the nearest centroid according to the number of clusters needed. This procedure minimizes the distance difference between data points in the same cluster and the centroid. To obtain a more accurate centroid, Lloyd's algorithm with squared Euclidean distances is applied to calculate each k in the model. In addition, the optimal number of clusters is being computed through Calinski-Harabasz criterion where the formula implied is:

$$\frac{SS_B}{SS_W} \times \frac{(N-k)}{(k-1)}$$

SS_B: between-cluster variance

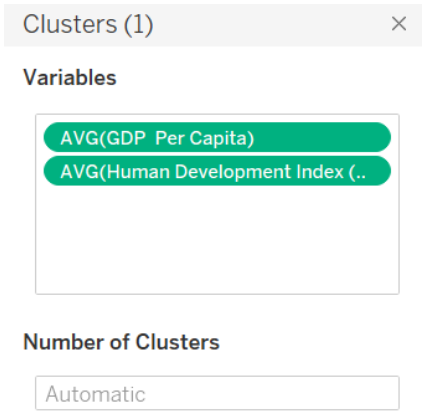
SS_W: within-cluster variance

k: number of clusters

N: number of observations

The advantage of clustering is the automated computation by the system to identify the groups of data which works well in numeric data. However, clustering cannot be applied on text-based data value and not suitable in cube data source as the variable inputs (Tableau, 2021).

Implementation



Result

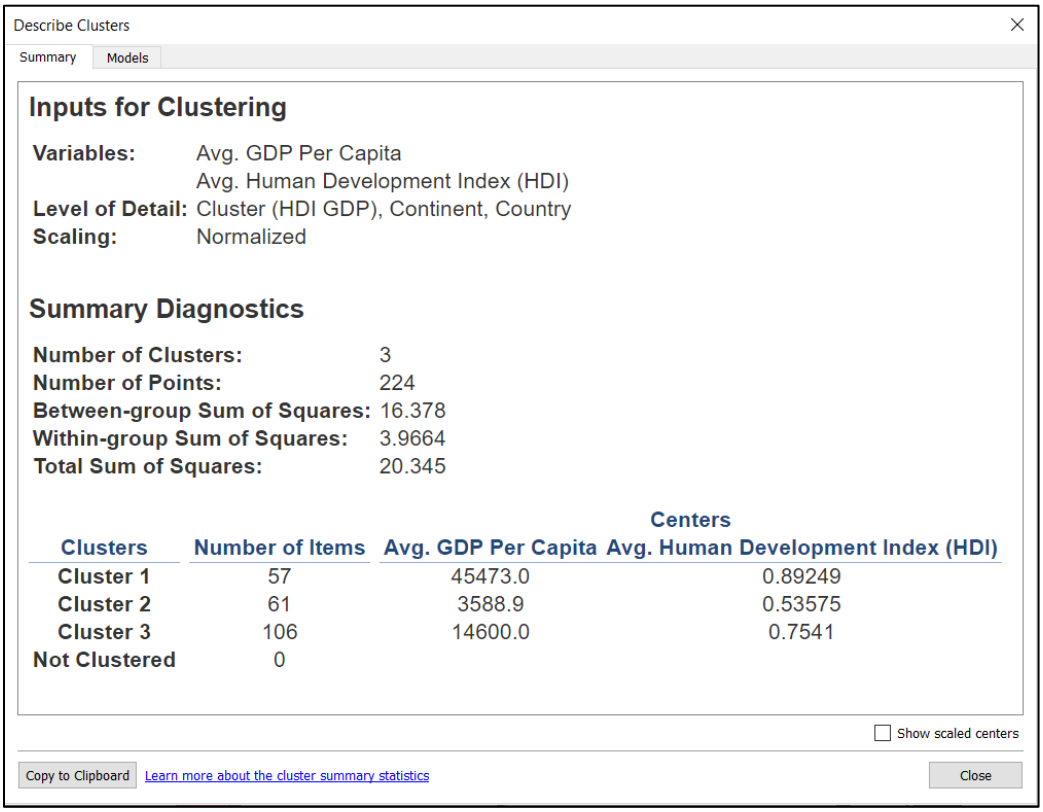


Figure 7.30: Clustering Summary

In the covid-economic corresponding map, there are three k in the k-mean clustering model. The variables that used to determine the clusters are average of GDP per capita and average human development index (HDI). Level of details that describe the data values are cluster group, continent name and country name. The model implements min-max normalization scaling where data values from every variable mapped to value between 0 and 1which minus the minimum of the mapped values and divide by the range of the mapped values. Calculation is as shown:

$$x – \min(x))/(\max(x) - \min(x)$$

From the diagnostics summary, the clustering model has a high between-group Sum of Squares indicating great separation between clusters. The low within-group Sum of Squares shows the cohesiveness of the data where the data fits well in the model. The large total sum of squares proves the well model quality (Tableau, 2021)

Analysis

The clusters are identified as categorized with meaningful name as listed:

- Cluster 1 (57 countries): High GDP Per Capita, High HDI
- Cluster 2 (61 countries): Low GDP Per Capita, Low HDI
- Cluster 3 (106 countries): Medium GDP Per Capita, Medium HDI

Generally, the clustering notices that the data with matching GDP and HDI in a parallel growth pattern. This eventually means that a country that has low GDP will has a low HDI as well. Logically, this statement is true because when a country has a great production on goods and services which relates to high country income. The country income directly raising the living standard and life quality of the citizens where the HDI will increase as well. Selecting a country from cluster 1 as an example, Australia (Oceania) has high standard of living which can be determine through the cost of living with the existence of great job prospects and professional financial services (Degree Plus Australia, 2021). The greatness in the human development, Australia has a high life expectation and well-being socioeconomic where the countries in other

country should learn from Australia for better economic growth in the GDP. For example, abides certain open market policies practiced by Australia including focus are on high trade freedom, investment freedom and financial freedom (The Heritage Foundation, 2021)

Analysis of Variance:

Variable	F-statistic	p-value	Model Sum of Squares	Model DF	Error Sum of Squares	Error DF
Avg. Human Development Index (HDI)	94.98	0.0	12.2	2	14.2	221
Avg. GDP Per Capita	75.05	0.0	4.175	2	6.146	221

Figure 7.31: Clustering Models Analysis of Variance (ANOVA)

The ANOVA table displays relevant analysis of the variables' variation within and between observations. Both HDI and GDP Per capita have a high F-statistic that indicates the well distinguish of corresponding variable in the clusters. The p-values of both variables are very low with a 0.0 explaining the expected values of the corresponding variable is independently differ among the clusters. In both variables, the Model Sum of Squares DF is small showing that the cluster means are close to the overall mean. Error Sum of Squares is the ratio of within-group sum of squares and error degrees of freedom.

7.3 Predictive analysis

In data mining, predictive analysis is being implemented to suggest better future financial allocation for each country. Supervised learning involved are trend line model (regression) and forecasting, meanwhile unsupervised learning conducted is clustering.

Task 1: Trend Line Analysis (Regression)

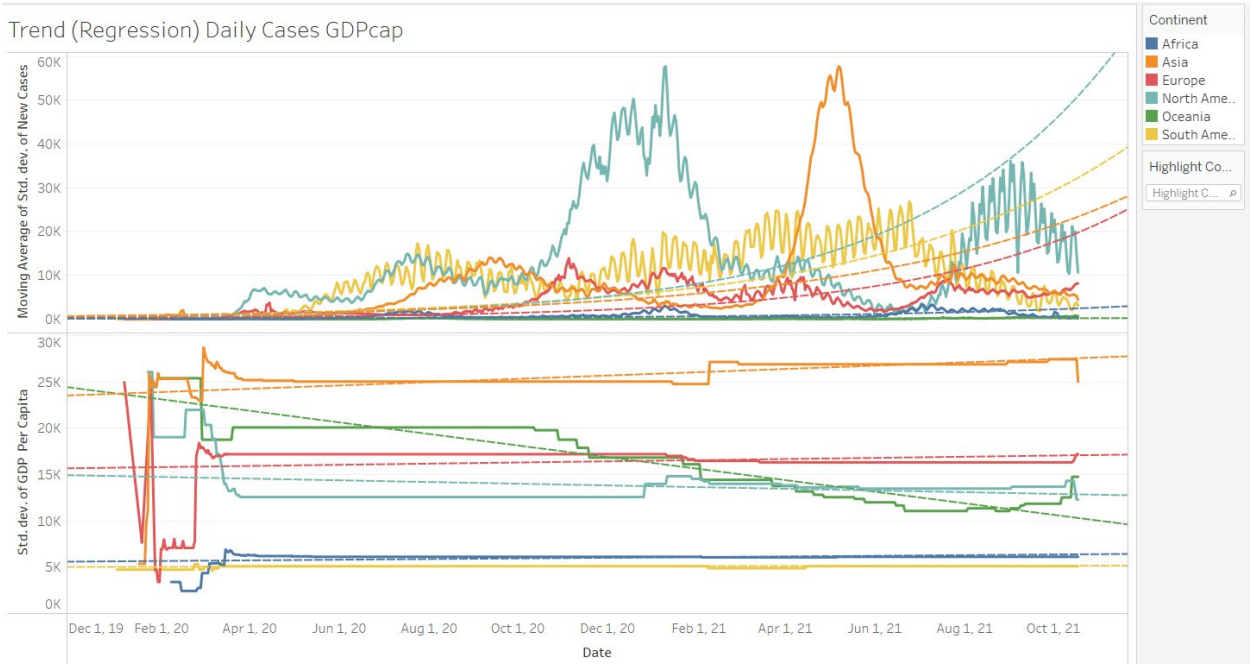


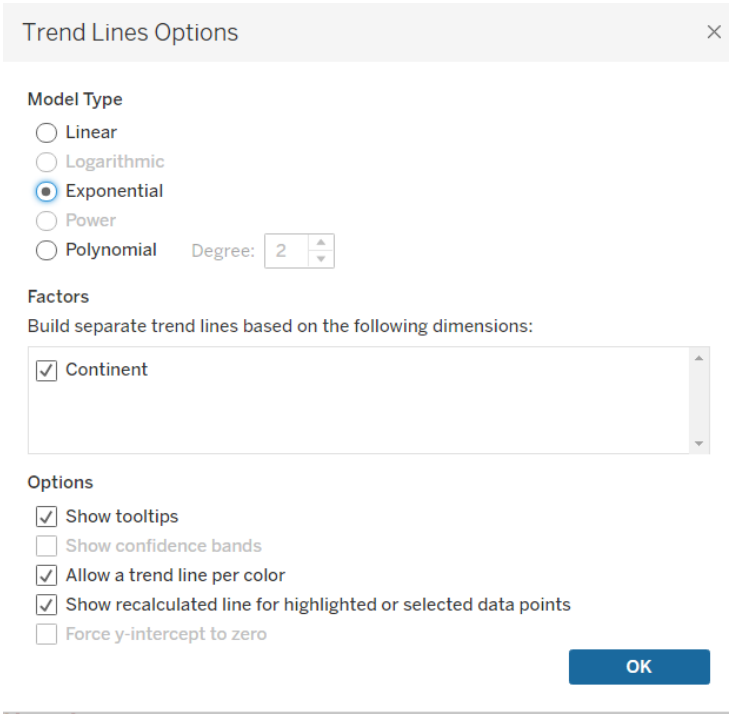
Figure 7.32: Trend Line Model (Regression) on Cases Trend, GDP Per Capita By Date

The trend line model is predicted through the regression algorithm. Basically, the regression algorithm predicts target numeric value (predictors) dependent on continuous data values. For example, the upper graph applies exponential regression that predict the growth and changes of continents' Covid new cases overtime. Meanwhile, the lower graph shows the application of linear regression in predicting GDP per capita overtime for each continent.

Note:

- The new cases values are being applied with calculation on moving average to smoothen the fluctuated data view.

Implementation



Linear

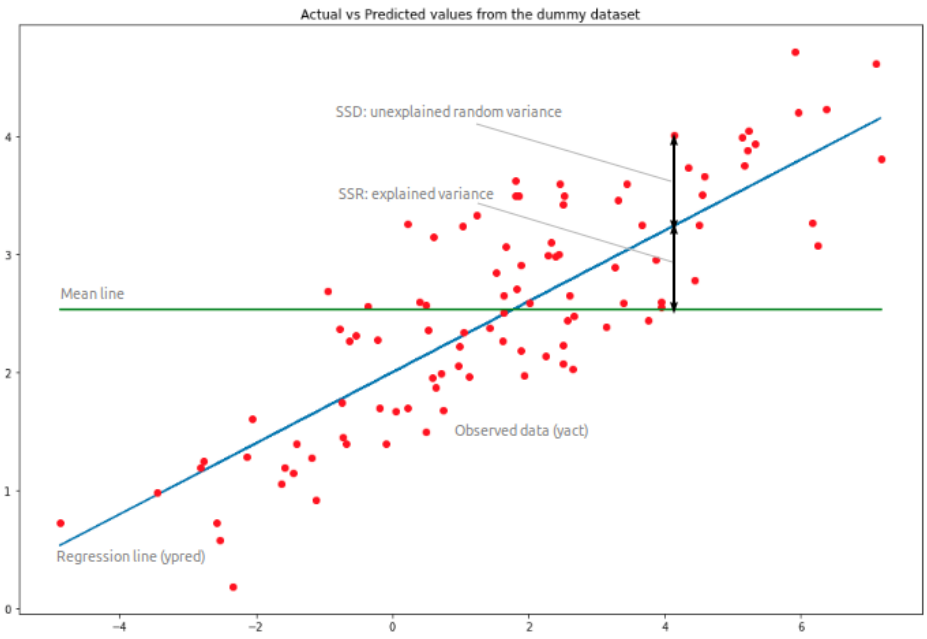


Figure 7.33: Linear Regression Example
(Li, 2018)

Linear regression is a common algorithm that estimates relationship between one target variable - dependent variable and one / more causing variable - independent variable (Reddy, 2020). The formula applied in this algorithm is (Tableau, 2021):

$$y = a_0 + a_1 * x$$

y = dependent variable

x = independent variable

a_0 = intercept

a_1 = slope

One of the pros of performing linear regression in a predictive analysis is the adaptability towards vast amount of dataset size. Moreover, linear regression helps analysts to determine the relevance of the selected features / variables. On the other hand, linear regression is exposed to the limitation of the weak assumption of the algorithm since the calculation is preferably simple where it is not applicable to complicated data trend (Reddy, 2020).

Result

Trend Lines Model

A linear trend model is computed for standard deviation of GDP Per Capita given Date. The model may be significant at p <= 0.05. The factor Continent may be significant at p <= 0.05.

Model formula:	Continent*(Date + intercept)
Number of modeled observations:	3811
Number of filtered observations:	0
Model degrees of freedom:	12
Residual degrees of freedom (DF):	3799
SSE (sum squared error):	9.58693e+09
MSE (mean squared error):	2.52354e+06
R-Squared:	0.953661
Standard error:	1588.57
p-value (significance):	< 0.0001

Figure 7.34: Trend Line Model (Linear Regression)

The linear trend line model predicts the standard deviation of GDP Per Capita with the formula of:

Stdev GDP Per Capita = (Date + intercept) * Continent

Through the observation on a total of 3811 data values, the linear algorithm identifies 12 model degrees of freedom as important parameters for the regression calculations with a residual

of 3799 values (observation number. – parameters estimated). The model has a high SSE (sum squared error) and MSE (mean squared error) where individually these values does not bring any meaning, but they are being further calculated for more impactful information. The R-squared (ratio of model's error variance) of the model is 95% plus which explains that the data fits well in the model and indicate the high model confidence. The R-squared computation is as shown:

$$1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Meanwhile, standard error of the model is high showing that the sample means are widely spread around the data. The model has a significance p-value of lesser than 0.05 (less) which indicates that the model is very important (Tableau, 2021).

Analysis

Analysis of Variance:

Field	DF	SSE	MSE	F	p-value
Continent	10	1.9669956e+11	1.967e+10	7794.59	< 0.0001

Figure 7.35: Trend Line (Linear) ANOVA

The ANOVA table displays all factors in the trend line model where the values are a compared version without furthering into the entire model.

Row	Column	Continent	p-value	DF	Term	Value	StdErr	t-value	p-value
Std. dev. of GDP Per Capita	Date	South America	< 0.0001	654	Date	0.208062	0.0241366	8.6202	< 0.0001
					intercept	-4132.95	1065.84	-3.87763	0.0001161
Std. dev. of GDP Per Capita	Date	Oceania	< 0.0001	625	Date	-20.5123	0.308523	-66.4854	< 0.0001
					intercept	922789	13628.5	67.7102	< 0.0001
Std. dev. of GDP Per Capita	Date	North America	< 0.0001	633	Date	-3.00679	0.445568	-6.74822	< 0.0001
					intercept	146589	19680.4	7.44847	< 0.0001
Std. dev. of GDP Per Capita	Date	Europe	< 0.0001	631	Date	2.04892	0.4739	4.32353	< 0.0001
					intercept	-74099.9	20932.3	-3.53997	0.0004296
Std. dev. of GDP Per Capita	Date	Asia	< 0.0001	639	Date	5.85922	0.418249	14.0089	< 0.0001
					intercept	-233118	18472.5	-12.6197	< 0.0001
Std. dev. of GDP Per Capita	Date	Africa	< 0.0001	617	Date	1.15616	0.135359	8.54146	< 0.0001
					intercept	-45086.2	5979.81	-7.53974	< 0.0001

Figure 7.36: Individual trend lines

Each trend line is being analysed with coefficient statistics. The value and term indicate the intercept value for each data variable axis (terms). From the table, each line is significantly

important based on the first p-value. The StdErr is defines as a measure on the spread of sampling distribution of the estimated coefficient. While t-value test the null hypothesis for true value on zero coefficient where the final p-value computes the probability of t-value to be larger in magnitude where the coefficient value is zero. Basically, the trend line model shows that Asia have a steadily increase GDP Per Capita and is expect to uprising more in the future. There are slight increasing trend for Europe and Africa continent where the GDP is predicted to have a slight escalate in the future. South America continent is having a prediction where GDP Per Capita will most likely remain as the past. However, for Oceania and North America might face a decline in GDP Per Capita in the future.

Exponential

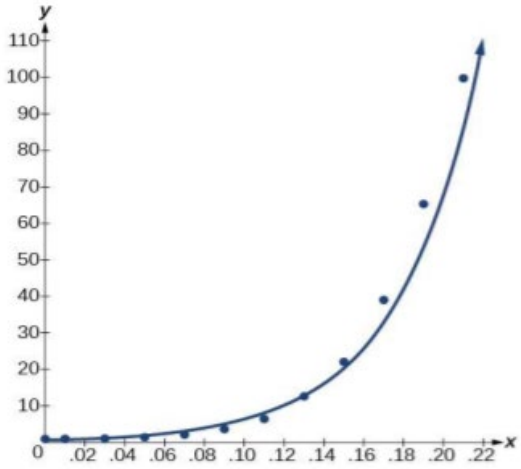


Figure 7.37: Exponential Regression Example
(Iqbal, 2020)

Exponential regression is a model algorithm that commonly used in estimating rapid growth and decay in the data. The algorithm will discover the best fits for the set of data and mostly will present as a curve line where the formula is as shown below:

$$y = \exp(b_0) * \exp(b_1 * x)$$

y: dependent variable

x: independent variable

$$\exp(b_0) > 0$$

The pro of the exponential is the accuracy on the prediction especially when there is certain decay in the data. However, the algorithm often predicts lagging trend which neglects the ups and downs of the data values. (Iqbal, 2020).

Result

A linear trend model is computed for natural log of Moving Average of Std. dev. of New Cases given Date. The model may be significant at p <= 0.05. The factor Continent may be significant at p <= 0.05.

Model formula:	Continent*(Date + intercept)
Number of modeled observations:	3711
Number of filtered observations:	100
Model degrees of freedom:	12
Residual degrees of freedom (DF):	3699
SSE (sum squared error):	9788.64
MSE (mean squared error):	2.64629
R-Squared:	0.648332
Standard error:	1.62674
p-value (significance):	< 0.0001

Figure 7.38: Trend Line Model (Exponential Regression)

The exponential trend line model predicts the standard deviation of GDP Per Capita with the formula of:

Moving Average Stdev New Cases = (Date + intercept) * Continent

Through the observation on a total of 3711 data values where 100 observations are ,being filtered the linear algorithm identifies 12 model degrees of freedom as important parameters for the regression calculations with a residual of 3699 values (observation number. – parameters estimated). The model has a high SSE (sum squared error) and low MSE (mean squared error) where individually these values does not bring any meaning, but they are being further calculated for more impactful information. The R-squared (ratio of model's error variance) of the model is 64.8% plus which explains that the data fits quite good in the model and indicate the moderate upper model confidence. The R-squared computation is as shown:

$$1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Meanwhile, standard error of the model is low showing that the sample means are closely placed around the data. The model has a significance p-value of lesser than 0.0001 (less) which indicates that the model is very important (Tableau, 2021)

Analysis

Analysis of Variance:

Field	DF	SSE	MSE	F	p-value
Continent	10	14031.492	1403.15	530.232	< 0.0001

Individual trend lines:

Panes		Color	Line		Coefficients				
Row	Column	Continent	p-value	DF	Term	Value	StdErr	t-value	p-value
Moving Average of Std. dev. of New Cases	Date	South America	< 0.0001	601	Date	0.005819	0.0004023	14.463	< 0.0001
					intercept	-248.488	17.7775	-13.9777	< 0.0001
Moving Average of Std. dev. of New Cases	Date	Oceania	< 0.0001	607	Date	0.0047023	0.0003434	13.6917	< 0.0001
					intercept	-204.499	15.1738	-13.4771	< 0.0001
Moving Average of Std. dev. of New Cases	Date	North America	< 0.0001	628	Date	0.0071417	0.0004924	14.5049	< 0.0001
					intercept	-306.87	21.7486	-14.1099	< 0.0001
Moving Average of Std. dev. of New Cases	Date	Europe	< 0.0001	629	Date	0.0070538	0.000359	19.6461	< 0.0001
					intercept	-303.909	15.8594	-19.1627	< 0.0001
Moving Average of Std. dev. of New Cases	Date	Asia	< 0.0001	632	Date	0.0054097	0.0002152	25.1399	< 0.0001
					intercept	-230.603	9.50473	-24.2619	< 0.0001
Moving Average of Std. dev. of New Cases	Date	Africa	< 0.0001	602	Date	0.0057284	0.0003234	17.7113	< 0.0001
					intercept	-247.077	14.2908	-17.2892	< 0.0001

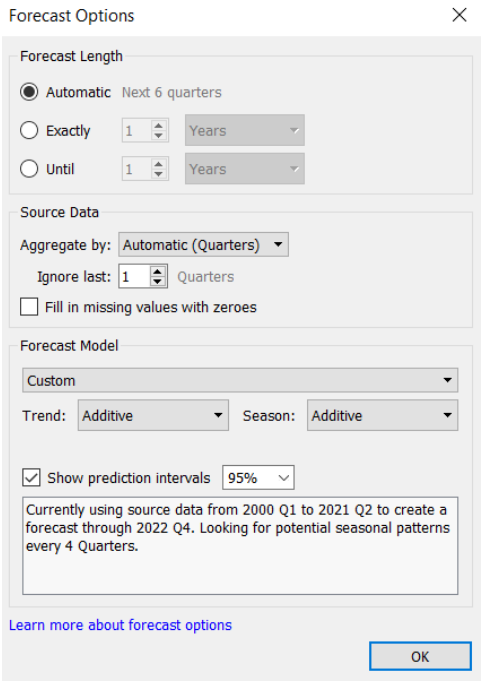
Figure 7.39: ANOVA and Individual Trend Lines

The ANOVA table displays all factors in the trend line model where the values are a compared version without furthering diving into the entire model. In the individual trend lines table, each trend line is being analysed with coefficient statistics. The value and term indicate the intercept value for each data variable axis (terms). From the table, each line is significantly important based on the small first p-value. The StdErr is defines as a measure on the spread of sampling distribution of the estimated coefficient. While t-value test the null hypothesis for true value on zero coefficient where the final p-value computes the probability of t-value to be larger in magnitude where the coefficient value is zero. Generally, all continents will have cases increase trend except for Oceania where the cases in the continent will be most likely to remain as it is. However, from the data visualization, logically some of the continent should be facing decline on new cases. This condition might happen due to the 65% of confidence and the limitations of the exponential algorithm.

Task 2: Forecasting

Forecasting is a scientific prediction that applied real-world data generating process (DGP). To obtain a high-quality forecast outcome, the forecast will discover a simple pattern in the DGP that match the anomalies described by model reasonably. The forecast implies certain quality metrics to measure the suitability of model in matching the DGP. The algorithm used is exponential smoothing which repetitively predicts the future value of a regular time series based on the weighted average of historical values. Furthermore, before any forecasting comes into place, the optimization on smoothing parameters. When there are not enough data to forecast, the model will forecast in getting a finer temporal granularity.

Implementation



With certain testing and data manipulation for forecasting, the best time horizon for the algorithm that is selected is quarters by year. This is to retrieve the best quality and accurate output from the forecasting algorithm. In the forecast model both additive and multiplicative are applied respectively to both line prediction and bar plot prediction. The additive model forecast through summed model components, multiplicative on the other hand forecast by having at least some components are being multiplied (Tableau, 2021).

Bar Chart

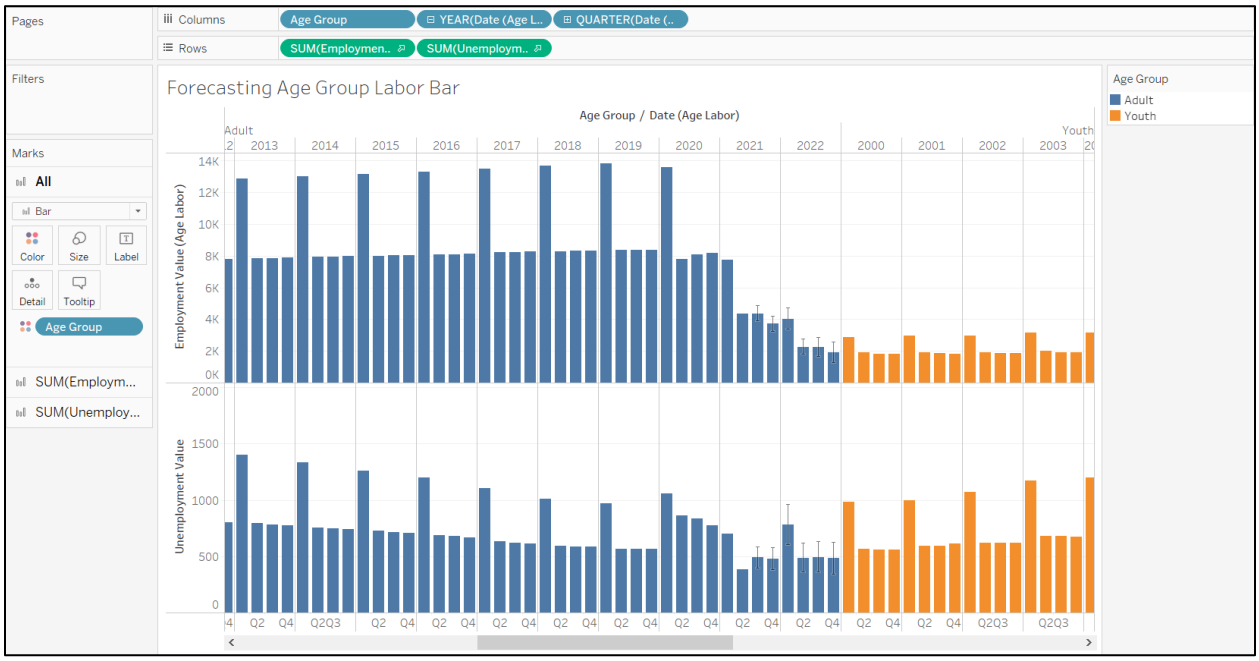


Figure 7.40: Age Group Labor Time Series Forecasting

The bar chart is used to forecast the sum of unemployment value and sum of employment values by age group (Adult & Youth).

Result

Options Used to Create Forecasts

Time series: Quarter of Date (Age Labor)
Measures: Sum of Employment Value (Age Labor), Sum of Unemployment Value
Forecast forward: 6 quarters (2021 Q3 – 2022 Q4)
Forecast based on: 2000 Q1 – 2021 Q2
Ignore last: 1 quarter (2021 Q3)
Seasonal pattern: 4 quarter cycle

Sum of Employment Value (Age Labor)

Column	Color	Initial	Change From Initial	Seasonal Effect		Contribution		
Age Group	Age Group	2021 Q3	2021 Q3 – 2022 Q4	High	Low	Trend	Season	Quality
Adult	Adult	4,358 ± 474	-2,443	2022 Q1 1	2022 Q2 1	100.0%	0.0%	Good
Youth	Youth	1,416 ± 138	-640	2022 Q1 1	2022 Q2 1	100.0%	0.0%	Good

Sum of Unemployment Value

Column	Color	Initial	Change From Initial	Seasonal Effect		Contribution		
Age Group	Age Group	2021 Q3	2021 Q3 – 2022 Q4	High	Low	Trend	Season	Quality
Adult	Adult	490 ± 93	-6	2022 Q1 1	2022 Q4 1	99.2%	0.8%	Good
Youth	Youth	734 ± 125	-33	2022 Q1 1	2022 Q2 1	99.9%	0.1%	Good

Figure 7.41: Forecast Model Summary

As mentioned, the selected time series horizon is the quarter of the date in Age Labor dataset. The data that is being measured / forecast are sum of employment value and sum of unemployment value. Forecast forward indicates the number of future predictions where 6 future quarters are being forecasted from 2021 Quarter 3 to 2022 Quarter 4. The forecasting is being predicted by using the existing data from 2000 Quarter 1 to 2021 Quarter 2 where the 2021 Q3 is being ignored because the existing data does not complete the whole quarter. The forecast is being implemented through a 4-quarter cycle seasonal pattern. The summary table describes each dimension (employment value & unemployment value) with measure details. Statistically, the initial describes the value and prediction interval for the first forecasted period. The seasonal effect displayed identify the model with iterative anomalies of variation overtime. The high low indication in seasonal effect where seasonal component expresses deviation of the trend. Contribution expresses the contribution of forecasting to either trend or season where from the diagram most of the contribution are being divided into trend. The quality of each forecast is in good standard which explains that the model fits in the data and the prediction is less likely to be a wrong value.

Analysis

All forecasts were computed using exponential smoothing.

Sum of Employment Value (Age Labor)

Column	Color	Model			Quality Metrics					Smoothing Coefficients		
Age Group	Age Group	Level	Trend	Season	RMSE	MAE	MASE	MAPE	AIC	Alpha	Beta	Gamma
Adult	Adult	Multiplicative	Multiplicative	Multiplicative	644	202	0.09	2.5%	1,130	0.500	0.500	0.500
Youth	Youth	Multiplicative	Multiplicative	Multiplicative	160	68	0.10	2.6%	891	0.455	0.500	0.500

Sum of Unemployment Value

Column	Color	Model			Quality Metrics					Smoothing Coefficients		
Age Group	Age Group	Level	Trend	Season	RMSE	MAE	MASE	MAPE	AIC	Alpha	Beta	Gamma
Adult	Adult	Multiplicative	Multiplicative	Multiplicative	94	45	0.20	6.0%	799	0.500	0.014	0.135
Youth	Youth	Multiplicative	Multiplicative	Multiplicative	122	57	0.19	5.4%	845	0.500	0.030	0.221

Figure 7.42: Forecast Models Details

The model that is being used for level, trend and season are all multiplicative. The quality metrics identified are RMSE (Root mean squared error), MAE (Mean absolute error), MASE (Mean absolute scaled error), MAPE (Mean absolute percentage error) and AIC (Akaike information criterion). The formulas of calculations are as displayed:

Basic formula residual:

$$e(t) = F(t)-A(t)$$

A(t): Actual value of the period – t

F(t): Forecast value at period – t

t: index of time series period

n: length of time series period

Metric	Formula
RMSE	$\sqrt{\left(\frac{1}{n}\right)\sum e(t)^2}$
MAE	$\frac{1}{n}\sum\left\|e(t)\right\|$
MASE - Measures magnitude of error compared to naive one-step ahead forecast magnitude of error	$\frac{\frac{1}{n}\sum\left\|e(t)\right\|}{\frac{1}{(n-1)}\sum_{2}^{n}\left\|Y(t)-Y(t-1)\right\|}$
MAPE - Measures magnitude of error compared to data magnitude (%)	$100\frac{1}{n}\sum\left\|\frac{e(t)}{A(t)}\right\|$
AIC - Model quality measure, avoid data overfitting	$n * \log(SSE/n) + 2 * (k+1)$

The smoothing coefficients are optimized by weighting the recent data over older values. Each coefficient place values indicate different smoothing coefficients as listed:

Alpha: Level smoothing coefficient

Beta: Trend smoothing coefficient

Gamma: Seasonal smoothing coefficient

Most of the smoothing coefficient in the table is far from 0 which indicates less smoothing is being performed (Tableau, 2021). Explaining the model prediction in a layman term, the employment value of adult is forecasted to be decreasing in the future with a slight lift in 2022 Quarter1. The unemployment of adult is forecasted with the pattern that is alike to the employment value but with higher values where 2022 Q2 – Q4 are in close value.

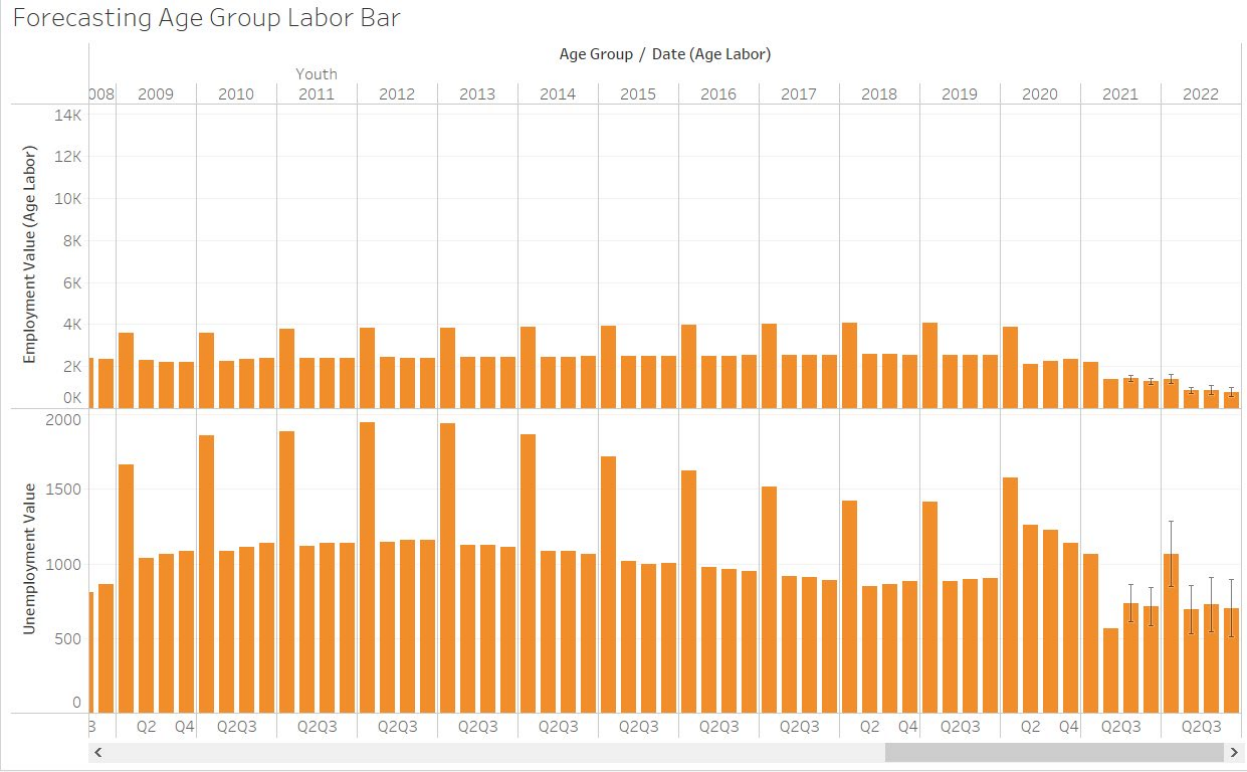


Figure 7.43: Youth Employ, Unemploy Forecast Result

The pattern forecasted in adult age group is most likely applied on the youth age group as well. The preferably clear difference is that the forecasted employment value for the Youth is very low compared to the adult.

Line Graph

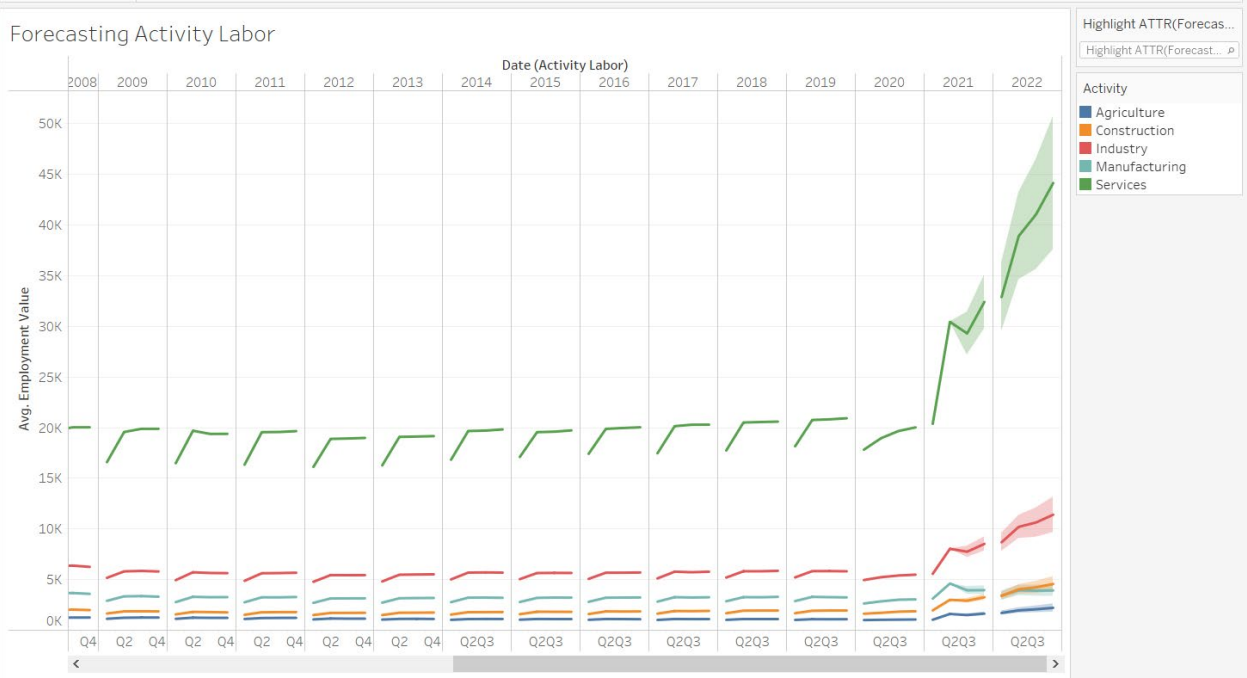


Figure 7.44: Activity Labor Time Series Forecasting

The line graph of average employment value grouped by types of activity is being forecasted for six future quarter in 2021 and 2022. Most of the activities are forecasted with an increasing trend in the future only the manufacturing sector is facing a decline based on the diagram.

Result

Options Used to Create Forecasts

Time series: Quarter of Date (Activity Labor)
Measures: Avg. Employment Value
Forecast forward: 6 quarters (2021 Q3 – 2022 Q4)
Forecast based on: 2000 Q1 – 2021 Q2
Ignore last: 1 quarter (2021 Q3)
Seasonal pattern: 4 quarter cycle

Avg. Employment Value

Color Activity	Initial 2021 Q3	Change From Initial 2021 Q3 – 2022 Q4	Seasonal Effect High		Seasonal Effect Low		Contribution Trend	Contribution Season	Quality
Services	29,300 ± 2,119	14,848	2022 Q2	1,510	2022 Q1	-1,585	96.3%	3.7%	Good
Manufacturing	3,940 ± 388	-11	2022 Q4	138	2022 Q1	-395	0.5%	99.5%	Good
Industry	7,753 ± 561	3,650	2022 Q2	454	2022 Q1	-350	96.3%	3.7%	Good
Construction	2,919 ± 235	1,647	2022 Q2	225	2022 Q1	-109	97.1%	2.9%	Good
Agriculture	1,497 ± 132	716	2022 Q2	57	2022 Q1	-54	98.2%	1.8%	Ok

Figure 7.45: Forecast Model Summary

The continuous time data selected is quarter of date in the Activity Labor dataset. The data that is being measured / forecast are average of employment value. Forecast forward is consist of 6 future quarters are being forecasted from 2021 Quarter 3 to 2022 Quarter 4. The forecasting is being predicted by using the existing data from 2000 Quarter 1 to 2021 Quarter 2 where the 2021 Q3 is being ignored because the existing data does not complete the whole quarter. The forecast is being implemented through a 4-quarter cycle seasonal pattern. The summary table describes each dimension (employment value) with measure details.

Analysis

All forecasts were computed using exponential smoothing.

Avg. Employment Value

Color Activity	Model Level	Model Trend	Model Season	Quality Metrics RMSE	MAE	MASE	MAPE	AIC	Smoothing Coefficients Alpha	Beta	Gamma
Services	Additive	Additive	Additive	1,081	373	0.22	1.8%	1,220	0.500	0.500	0.126
Manufacturing	Additive	Additive	Additive	198	98	0.34	2.9%	928	0.500	0.000	0.000
Industry	Additive	Additive	Additive	286	125	0.30	2.1%	991	0.500	0.500	0.173
Construction	Additive	Additive	Additive	120	50	0.29	2.5%	841	0.500	0.500	0.217
Agriculture	Additive	Additive	Additive	67	30	0.57	2.6%	742	0.500	0.500	0.080

Figure 7.46: Forecast Models Details

The model that is being used for level, trend and season are additive. The quality metrics applied are RMSE (Root mean squared error), MAE (Mean absolute error), MASE (Mean absolute scaled error), MAPE (Mean absolute percentage error) and AIC (Akaike information criterion).

The Alpha and Beta smoothing coefficient of each activity is far from 0.00 with a high value of 0.5. However, the seasonal smoothing coefficient (gamma) being performed with high data smoothing which beneficial in enabling gradual component changes and less dependency on recent data (Tableau, 2021). In summary, the Services activity is predicted to increase a lot in the future, while Industry, Agriculture & Construction will increase with a minimal change. Yet, the Manufacturing sector will decline based on the forecasted result.

8.0 Ethical Issues & Social Impacts

Despite the implementation in data analytics and data evaluation, data research is often prone to ethical issues. Firstly, the datasets obtained are all available online which then suggest the ethical problems on data privacy. Information theft might try to break through the security layer of the datasets to reach out the origin data source location and sell private data to the public. The open-source data might be edited by unauthorized personnel where the issue is being extended to data accuracy. Due to the exposure of the data, issues like value and risk towards certain individuals or countries are highlighted. From the inspection of the data, weak countries identified through the economic indicators will most likely get targeted as a target victim by unethical personnel. Furthermore, the documentation and dissemination of fraud data is also an ethical issue for online datasets.

The social impacts of the entire research mainly contribute mainly to the economic sector. The suggestions on certain policies decision making, changing operation model and innovation on labor force is provided. From the perspective of resource planning, evaluation on distributing and managing the commodities are recommended for better costs and benefits for countries to implement financial allocation for countries growth. To recover from the economic crisis, the research report suggests the countries to better maintaining policies and focus on the business reoperation for safe areas. The consideration on human development is crucial for countries to recover from both pandemics and economic inflation due to Covid-19 virus outbreak.

9.0 Conclusion

In conclusion, the implementation of the Covid-19 data-driven research is conducted through KDD methodology. The research begins with the determination on domain, objective and scope of responsibilities by each team member. With a proposed idea, researcher conducts data selection for the scope that is desired to be achieved. The research is then continued with data cleaning, data transformation, data visualization and data mining. In terms of data evaluation, the countries are having better control towards the Covid-19 pandemic with the rising of vaccinations where business operation is starting to work as usual once again. Most of the countries face a decline in labour force and working hours in 2020 where the year is a peak year of Covid-19 virus outbreak. Since most of the countries are lifting certain policies with the increasing vaccinations, the business is starting to operate as usual once again. Through the inspection on data mining and visualization, the data patterns are being discovered and suggestions are provided to solve the countries that have certain issue identified from the data. With the recommendations given, the countries are expected to distribute and plan a better financial solution to recover from the economic crisis affected by the covid pandemic.

All files in implementing the research:
https://github.com/Laikaiyong/Covid-19-Economic-Impact-Analysis

10.0 Personal reflection report

Throughout the research and implementation, I feel delightful to have the chance to explore the basic of data analytics. The fundamental knowledge in data selection, data cleaning, data transformation, data visualization and data mining is essential to me, I am confident to say that the course module had provide me a strong base knowledge towards data analytics. As a data science enthusiast, I am happy to experience through such amazing and interesting subject. The technical experience on Tableau software helps me to get a brief idea on a data analyst's work basis. On the other hand, the requirement of documenting a report further my knowledge theoretically on data analysis.

To express some gratitude to my lecturer and teammates, I am glad to have such cooperative team members. With the guidance and approval from my lecturer, I get a better idea in working on a data driven research. This course subject ignites my passion in data related research, and I am planning to work on machine learning face recognition self-project in the semester break. The insights gain from the analysis and evaluation is beneficial to me and these tasks eventually train me to think like a real data analyst and drive decision or recommendation with a scientific data approach.

11.0 References

Ahmed, K. (2020, March 26). *Debt relief allows Somalia to rejoin global economy after 30-year exile*. Retrieved from The Guardian: https://www.theguardian.com/global-development/2020/mar/26/debt-relief-allows-somalia-to-rejoin-global-economy-after-30-year-exile

Bauer, L., Broady, K., Edelberg, W., & O'Donnel, J. (2020). *Ten facts about COVID-19 and the U.S. economy.* United States: Brookings.

Betti, F., & Heinzmann, T. (2020, March 24). *From perfume to hand sanitiser, TVs to face masks: how companies are changing track to fight COVID-19*. Retrieved from World Economic Forum: https://www.weforum.org/agenda/2020/03/from-perfume-to-hand-sanitiser-tvs-to-face-masks-how-companies-are-changing-track-to-fight-covid-19/

Blavatnik School of Government. (2021). *COVID-19 Government Response Tracker*. Retrieved from Blavatnik School of Government: https://www.bsg.ox.ac.uk/research/research-projects/covid-19-government-response-tracker

Callen, T. (2020, February 24). *Gross Domestic Product: An Economy's All*. Retrieved from International Monetary Fund: https://www.imf.org/external/pubs/ft/fandd/basics/gdp.htm

Chen, J. (2021, August 19). *Trendline*. Retrieved from Investopedia: https://www.investopedia.com/terms/t/trendline.asp

Chu, M. M. (2021, October 10). *Malaysia lifts travel restrictions for fully vaccinated people*. Retrieved from Reuters: https://www.reuters.com/world/asia-pacific/malaysia-lifts-travel-restrictions-fully-vaccinated-people-2021-10-10/

Degree Plus Australia. (2021). *Standard of Living*. Retrieved from Degree Plus Australia: https://pamdpa.com.au/why-australia/living-standards/

Gan, N., & Yeung, J. (2021, Septemver 29). *China has built a 5,000-room quarantine center for overseas arrivals. It could be the first of many*. Retrieved from CNN World: https://edition.cnn.com/2021/09/29/china/guangzhou-covid-quarantine-center-mic-intl-hnk/index.html

Goldman, D., & Tappe, A. (2021, November 12). *Americans haven't felt this bad about the economy in a decade*. Retrieved from CNN Business: https://edition.cnn.com/2021/11/12/economy/consumer-sentiment/index.html

Hamilton, H. (2018, July 9). *Overview of the KDD Process*. Retrieved from University of Regina: http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html

Harari, D., Keep, M., & Brien, P. (2021, September 24). *Coronavirus: Economic impact*. Retrieved from UK Parliament: https://commonslibrary.parliament.uk/research-briefings/cbp-8866/

IOM UN Migration. (2021). *Covid-19 Disease South America Regional Response.* Brazil: IOM UN Migration.

Iqbal, M. A. (2020). Application of Regression Techniques with their Advantages and Disadvantages. *Elektron Magazine*, 17.

Keane, J. (2020, December 15). *How The Pandemic Put Food Delivery Firms In The Limelight In 2020*. Retrieved from Forbes: https://www.forbes.com/sites/jonathankeane/2020/12/15/how-the-pandemic-put-food-delivery-firms-in-the-limelight-in-2020/?sh=1d1ec2ac5eeb

Kramer, L. (2021, June 17). *How Do Governments Reduce Inflation?* Retrieved from Investopedia: https://www.investopedia.com/ask/answers/111314/what-methods-can-government-use-control-inflation.asp

Li, L. (2018, October 25). *Introduction to Linear Regression in Python*. Retrieved from Towards Data Science: https://towardsdatascience.com/introduction-to-linear-regression-in-python-c12a072bedf0

McLeod, S. (2019). *What does a box plot tell you?* Retrieved from Simply Psychology: https://www.simplypsychology.org/boxplots.html

Miller, M. E. (2021, October 11). *Sydney starts to live with covid after 106-day lockdown. First stop: The pub.* Retrieved from The Washington Post: https://www.washingtonpost.com/world/asia_pacific/sydney-reopening-covid-lockdown/2021/10/11/538a8d20-1f3f-11ec-a8d9-0827a2a4b915_story.html

Nist/Sematech. (2012, April). *What are outliers in the data?* Retrieved from Engineering Statistics Handbook: https://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm

OLAP.com. (2021). *What is the definiyion of OLAP?* Retrieved from OLAP.com: https://olap.com/olap-definition/

Orlansky, D., & Boruchowicz, G. (2020, November 10). *Argentina: COVID-19 – Business Shutdowns & Reductions (Labor)*. Retrieved from Global Compliance News: https://www.globalcompliancenews.com/2020/11/10/argentina-business-shutdowns-and-reductions-in-force28092020/

Priy, S. (2021, September 22). *Clustering in Machine Learning*. Retrieved from Geeks for Geeks: https://www.geeksforgeeks.org/clustering-in-machine-learning/

Raghavan. (2018, August 16). *Scatter matrix , Covariance and Correlation Explained*. Retrieved from Medium: https://medium.com/@raghavan99o/scatter-matrix-covariance-and-correlation-explained-14921741ca56

Ray, S. (2015, August 14). *7 Regression Techniques you should know!* Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/

Reddy, E. (2020, July 17). *Advantages and Disadvantages of different Regression models*. Retrieved from Geeks for Geeks: https://www.geeksforgeeks.org/advantages-and-disadvantages-of-different-regression-models/

Sabri, I. A., Man, M., Abu Bakar, W. A., & Rose, A. N. (2019). Web Data Extraction Approach for Deep Web using WEIDJ. *Elsevier B.V.* , 11.

Stedman, C., & Burns, E. (2020, September). *business intelligence (BI)*. Retrieved from Search Business Analytics: https://searchbusinessanalytics.techtarget.com/definition/business-intelligence-BI

Sutner, S. (2020, August). *Business intelligence dashboard*. Retrieved from Search Business Analytics: https://searchbusinessanalytics.techtarget.com/definition/business-intelligence-dashboard

Tableau. (2021). *Add Trend Lines to a Visualization*. Retrieved from Tableaiu: https://help.tableau.com/current/pro/desktop/en-us/trendlines_add.htm

Tableau. (2021). *Data Types*. Retrieved from Tableau: https://help.tableau.com/current/pro/desktop/en-us/datafields_typesandroles_datatypes.htm

Tableau. (2021). *Find Clusters in Data*. Retrieved from Tableau: https://help.tableau.com/current/pro/desktop/en-us/clustering.htm

Tableau. (2021). *Forecast Descriptions*. Retrieved from Tableau: https://help.tableau.com/current/pro/desktop/en-us/forecast_describe.htm

Tableau. (2021). *How Forecasting Works in Tableau*. Retrieved from Tableau: https://help.tableau.com/current/pro/desktop/en-us/forecast_how_it_works.htm

Tableau. (2021). *Number Functions*. Retrieved from Tableau: https://help.tableau.com/current/pro/desktop/en-us/functions_functions_number.htm

Tableau. (2021). *Tableau Functions (by Category)*. Retrieved from Tableau: https://help.tableau.com/current/pro/desktop/en-us/functions_all_categories.htm

Tableau. (2021). *Time Series Forecasting: Definition, Applications, and Examples*. Retrieved from Tableau: https://www.tableau.com/learn/articles/time-series-forecasting

The Heritage Foundation. (2021). *Australia*. Retrieved from 2021 Index of Economic Freedom: https://www.heritage.org/index/country/australia

The Star. (2021, August 9). *Higher jobless rate in June due to tighter Covid-19 curbs*. Retrieved from The Star: https://www.thestar.com.my/business/business-news/2021/08/09/higher-jobless-rate-in-june-due-to-tighter-covid-19-curbs

UBS. (2021, March 16). *Mexico: COVID and the Impact on Women in the Labour Force*. Retrieved from UBS: https://www.ubs.com/global/en/investment-bank/in-focus/covid-19/2021/women-employment-rates.html

UNDP. (2021). *Human Development Index (HDI)*. Retrieved from United Nations Development Programme: http://hdr.undp.org/en/content/human-development-index-hdi

US News. (2020). *Colombia*. Retrieved from US News: https://www.usnews.com/news/best-countries/colombia

WHO Africa. (2021, October 28). *Less than 10% of African countries to hit key COVID-19 vaccination goal* . Retrieved from WHO Africa: https://www.afro.who.int/news/less-10-african-countries-hit-key-covid-19-vaccination-goal

Wolters Kluwer. (2021). *OLAP modeling and dashboards*. Retrieved from Wolters Kluwer: https://www.wolterskluwer.com/en/solutions/cch-tagetik/glossary/olap-modeling-and-dashboards