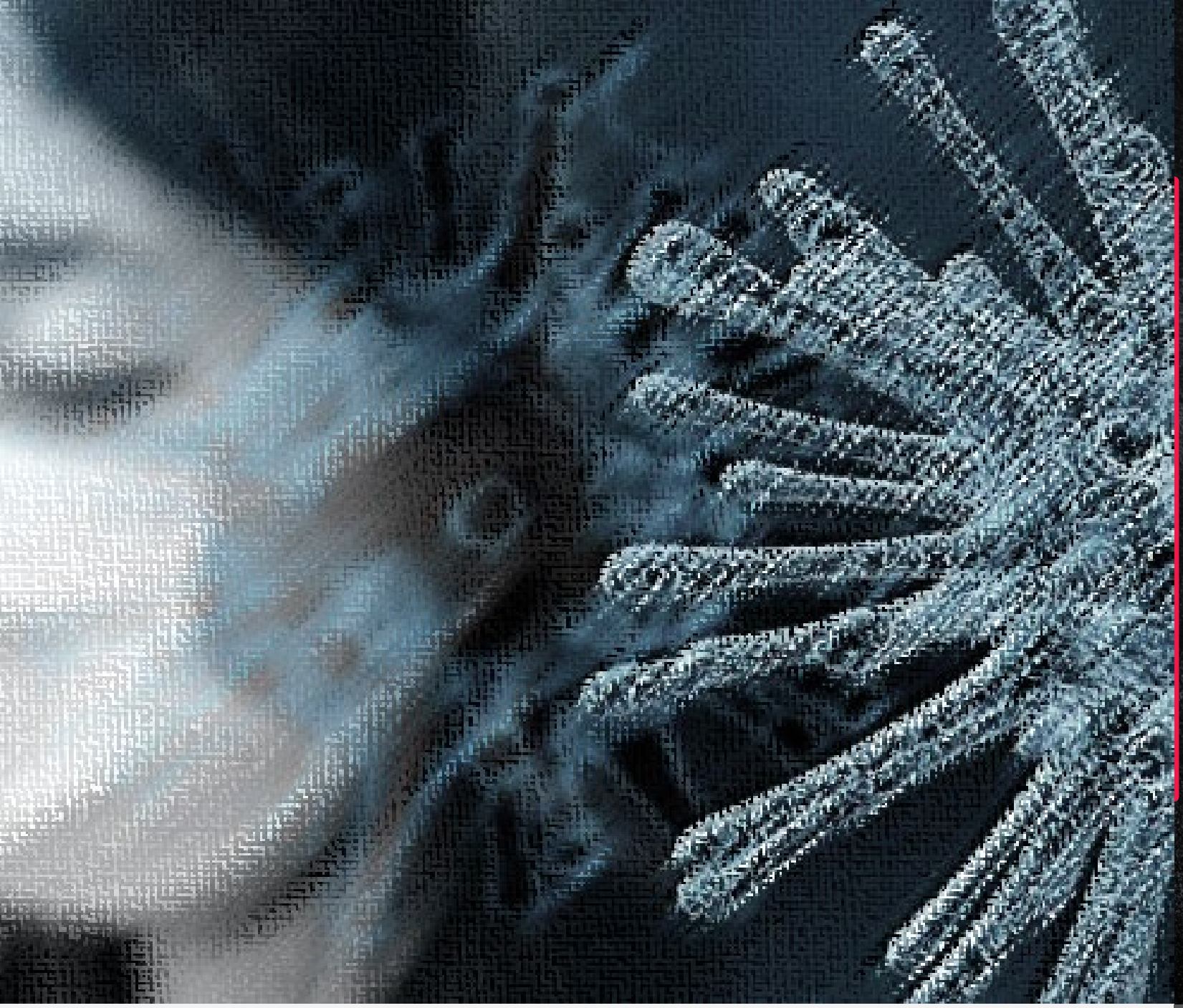


Combating COVID-19 Outbreaks

WQD7005 Data Mining
Master of Data Science | University of Malaya

Part D: Interpretation & Communication of Data Insights (Individual)

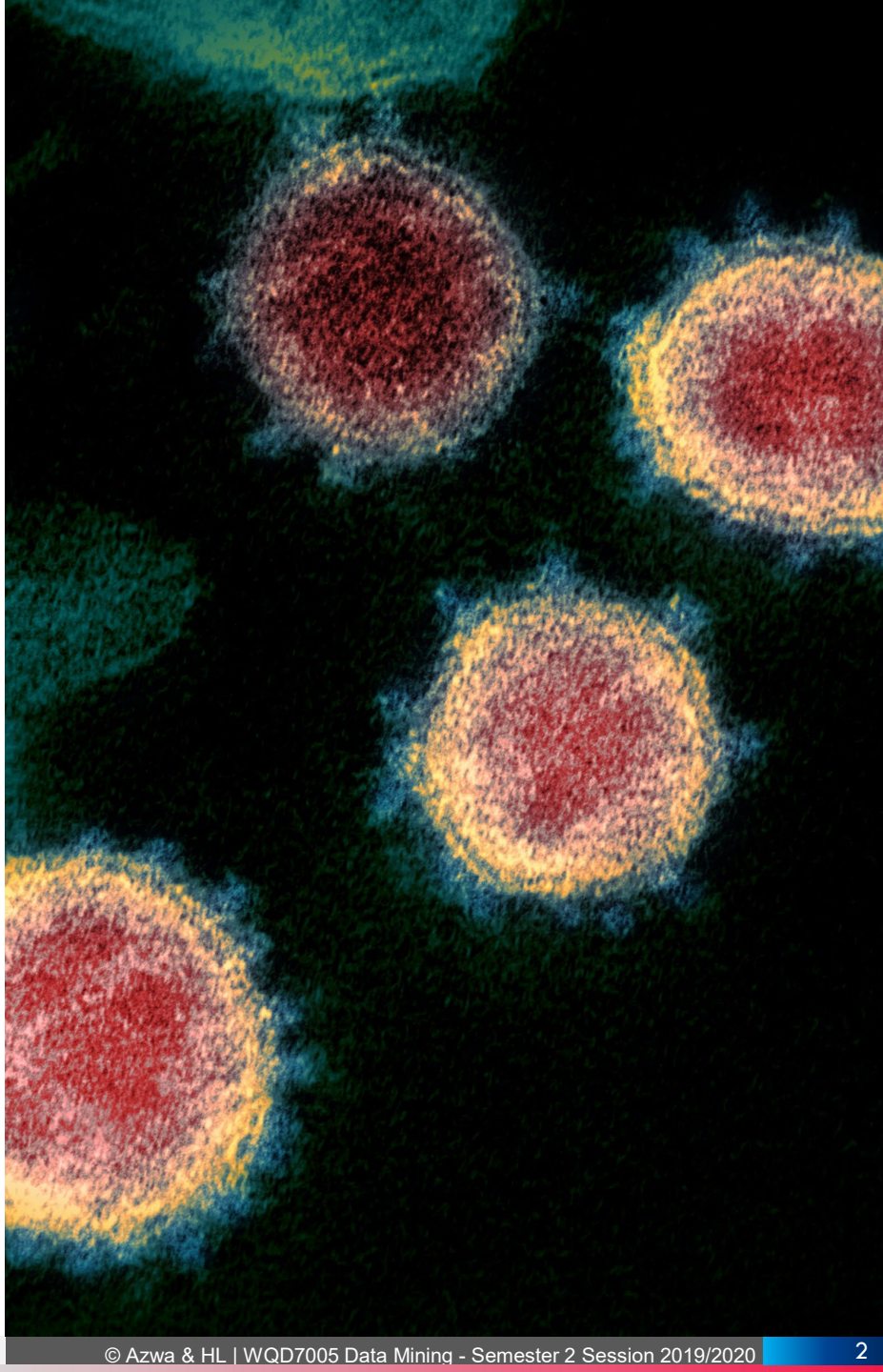
Student Name:
Kok Hon Loong (WQD170086)

UNIVERSITY OF MALAYA

Assignment Milestones

WQD7005 Data Mining | Semester 2 Session 2019/2020

- Part A: (Group)
  - Web Crawling of Real-time Data
- Part B: (Group)
  - Management of Data using Hadoop Data Warehouse or Data Lake
- Part C: (Group)
  - Accessing and Processing of Data from Hadoop Data Warehouse or Data Lake using Python
- **Part D: (Individual)**
  - **Interpretation and Communication of Data Insights**
- Part E: (Group)
  - Deployment of the Data Mining Results on Web (Flask) and Mobile Application (Kivy)



Assignment Background

Recent extraordinary improvements in data-collecting technologies have changed the way for the data scientist to make informed and effective decisions.

In this assignment, I will focus on the effective methods in interpreting and then communicating the analytics work from the data acquired. This includes:

- Methods for translating the data into information and possibly knowledge;
- Using statistical and visualization methods for creating artifacts to deliver the insights;
- Building compelling data presentations to communicate the findings to the stakeholders.

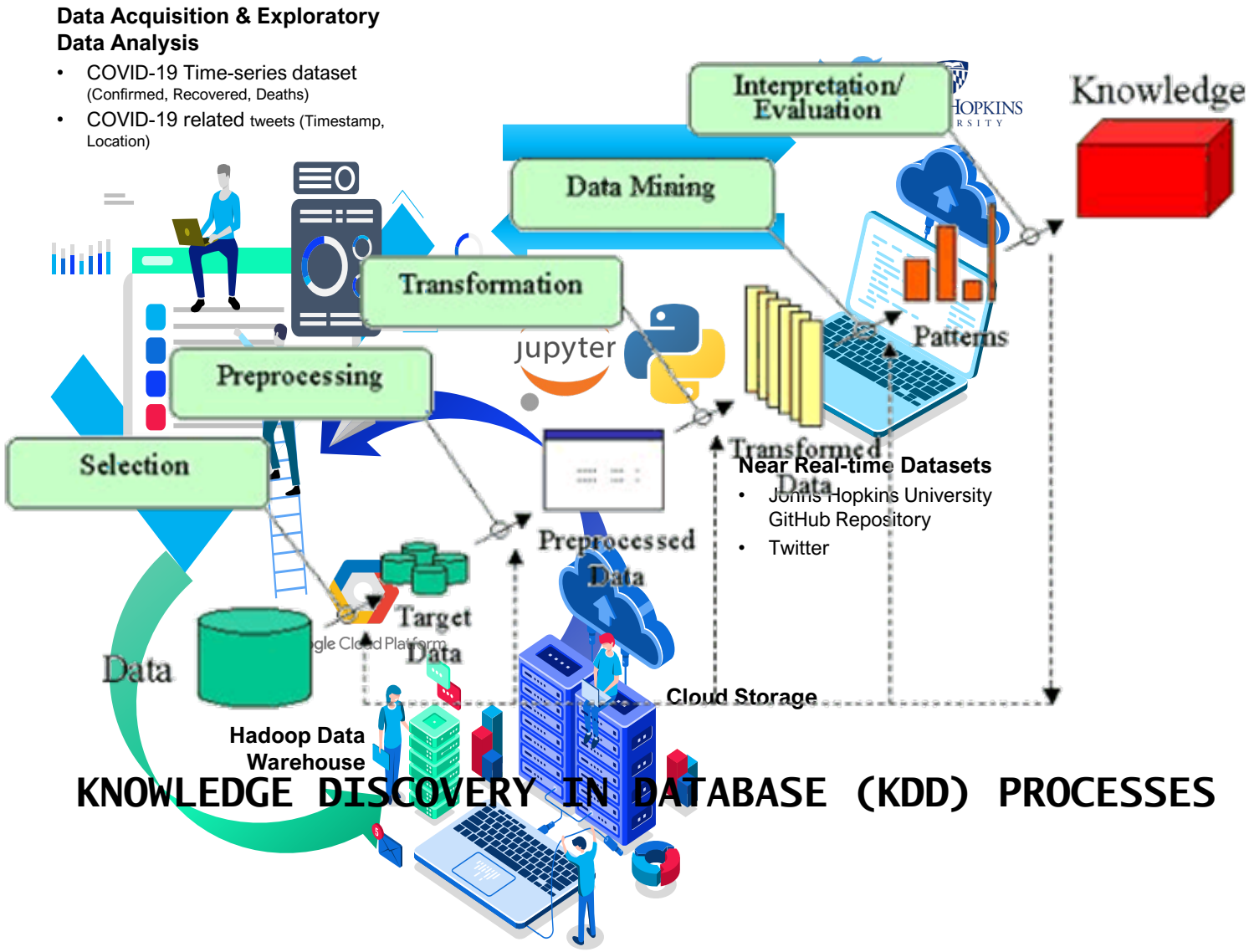
The Python code which was use as the tool to perform the analysis and forecasting on the data acquired is also uploaded in my group assignment GitHub at the link below:

- https://github.com/hlkok/WQD7005DataMining-Assignments

UNIVERSITY OF MALAYA

Process of Acquiring the Datasets

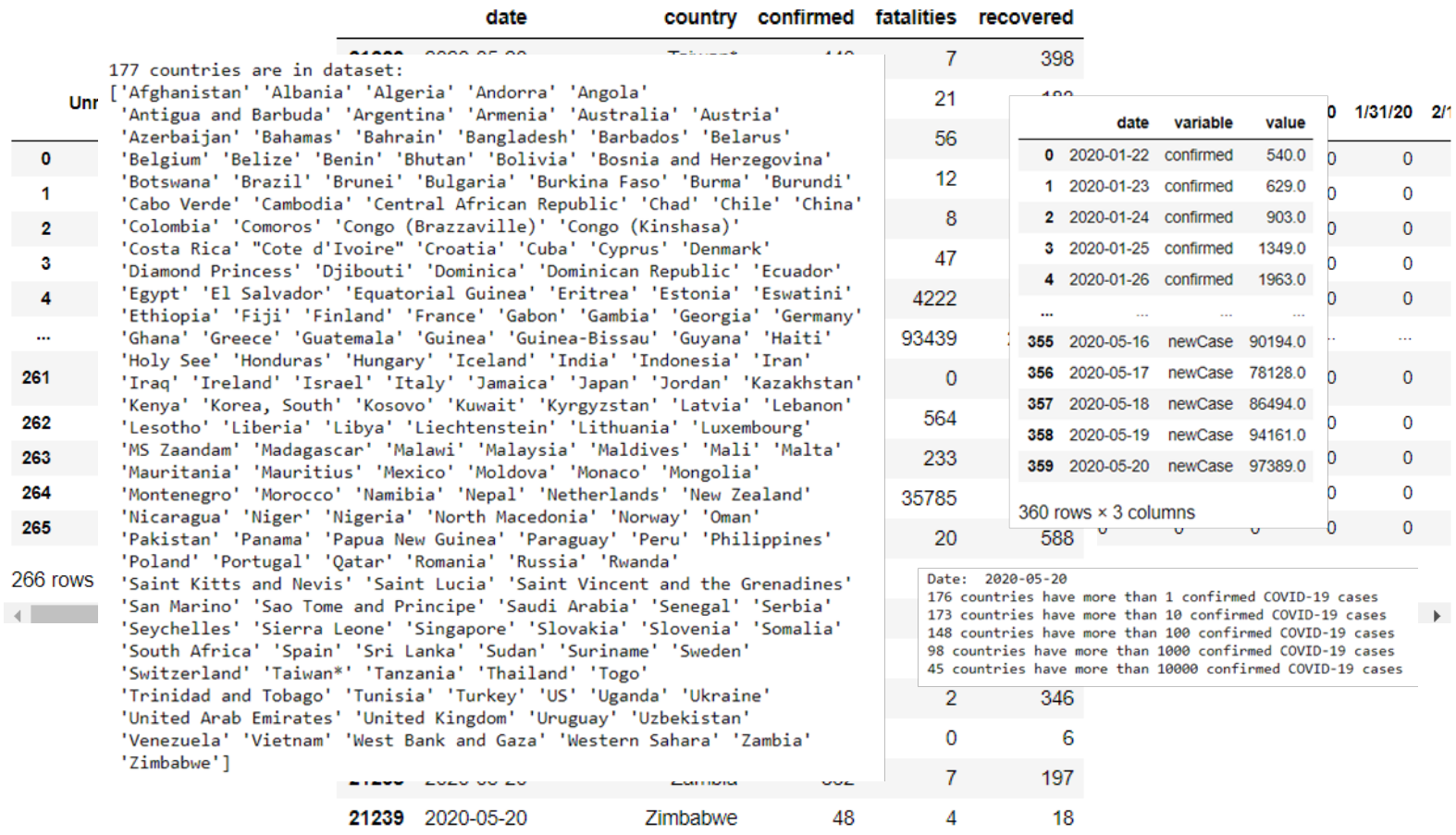
storing in Cloud Storage (Data Lake)



- The COVID-19 datasets were acquired based on web crawling of the near real-time data from Internet.
- Using the Data Lake to store the multiple data types acquired (structured and un-structured.
- Developed the visualization, and analytics forecasting tool using Python to present the data insights leveraging on the KDD processes as illustrated on the left.

COVID-19 Datasets

Preprocessing & Transformation of the Datasets



	date	country	confirmed	fatalities	recovered
				7	398
				21	
				56	
				12	
				8	
				47	
				4222	
				93439	
				0	
				564	
				233	
				35785	
				20	588
				2	346
				0	6
				7	197
21239	2020-05-20	Zimbabwe	48	4	18

```
177 countries are in dataset:
['Afghanistan' 'Albania' 'Algeria' 'Andorra' 'Angola'
 'Antigua and Barbuda' 'Argentina' 'Armenia' 'Australia' 'Austria'
 'Azerbaijan' 'Bahamas' 'Bahrain' 'Bangladesh' 'Barbados' 'Belarus'
 'Belgium' 'Belize' 'Benin' 'Bhutan' 'Bolivia' 'Bosnia and Herzegovina'
 'Botswana' 'Brazil' 'Brunei' 'Bulgaria' 'Burkina Faso' 'Burma' 'Burundi'
 'Cabo Verde' 'Cambodia' 'Central African Republic' 'Chad' 'Chile' 'China'
 'Colombia' 'Comoros' 'Congo (Brazzaville)' 'Congo (Kinshasa)'
 'Costa Rica' "Cote d'Ivoire" 'Croatia' 'Cuba' 'Cyprus' 'Denmark'
 'Diamond Princess' 'Djibouti' 'Dominica' 'Dominican Republic' 'Ecuador'
 'Egypt' 'El Salvador' 'Equatorial Guinea' 'Eritrea' 'Estonia' 'Eswatini'
 'Ethiopia' 'Fiji' 'Finland' 'France' 'Gabon' 'Gambia' 'Georgia' 'Germany'
 'Ghana' 'Greece' 'Guatemala' 'Guinea' 'Guinea-Bissau' 'Guyana' 'Haiti'
 'Holy See' 'Honduras' 'Hungary' 'Iceland' 'India' 'Indonesia' 'Iran'
 'Iraq' 'Ireland' 'Israel' 'Italy' 'Jamaica' 'Japan' 'Jordan' 'Kazakhstan'
 'Kenya' 'Korea, South' 'Kosovo' 'Kuwait' 'Kyrgyzstan' 'Latvia' 'Lebanon'
 'Lesotho' 'Liberia' 'Libya' 'Liechtenstein' 'Lithuania' 'Luxembourg'
 'MS Zaandam' 'Madagascar' 'Malawi' 'Malaysia' 'Maldives' 'Mali' 'Malta'
 'Mauritania' 'Mauritius' 'Mexico' 'Moldova' 'Monaco' 'Mongolia'
 'Montenegro' 'Morocco' 'Namibia' 'Nepal' 'Netherlands' 'New Zealand'
 'Nicaragua' 'Niger' 'Nigeria' 'North Macedonia' 'Norway' 'Oman'
 'Pakistan' 'Panama' 'Papua New Guinea' 'Paraguay' 'Peru' 'Philippines'
 'Poland' 'Portugal' 'Qatar' 'Romania' 'Russia' 'Rwanda'
 'Saint Kitts and Nevis' 'Saint Lucia' 'Saint Vincent and the Grenadines'
 'San Marino' 'Sao Tome and Principe' 'Saudi Arabia' 'Senegal' 'Serbia'
 'Seychelles' 'Sierra Leone' 'Singapore' 'Slovakia' 'Slovenia' 'Somalia'
 'South Africa' 'Spain' 'Sri Lanka' 'Sudan' 'Suriname' 'Sweden'
 'Switzerland' 'Taiwan*' 'Tanzania' 'Thailand' 'Togo'
 'Trinidad and Tobago' 'Tunisia' 'Turkey' 'US' 'Uganda' 'Ukraine'
 'United Arab Emirates' 'United Kingdom' 'Uruguay' 'Uzbekistan'
 'Venezuela' 'Vietnam' 'West Bank and Gaza' 'Western Sahara' 'Zambia'
 'Zimbabwe']
```

	date	variable	value
0	2020-01-22	confirmed	540.0
1	2020-01-23	confirmed	629.0
2	2020-01-24	confirmed	903.0
3	2020-01-25	confirmed	1349.0
4	2020-01-26	confirmed	1963.0
...	...	...	...
355	2020-05-16	newCase	90194.0
356	2020-05-17	newCase	78128.0
357	2020-05-18	newCase	86494.0
358	2020-05-19	newCase	94161.0
359	2020-05-20	newCase	97389.0

360 rows × 3 columns

```
Date:  2020-05-20
176 countries have more than 1 confirmed COVID-19 cases
173 countries have more than 10 confirmed COVID-19 cases
148 countries have more than 100 confirmed COVID-19 cases
98 countries have more than 1000 confirmed COVID-19 cases
45 countries have more than 10000 confirmed COVID-19 cases
```

- Removal of outliers, and apply strategies for handling missing data.
- Perform data reduction leveraging on useful features to represent the data.
- Collect necessary information to model for data mining.

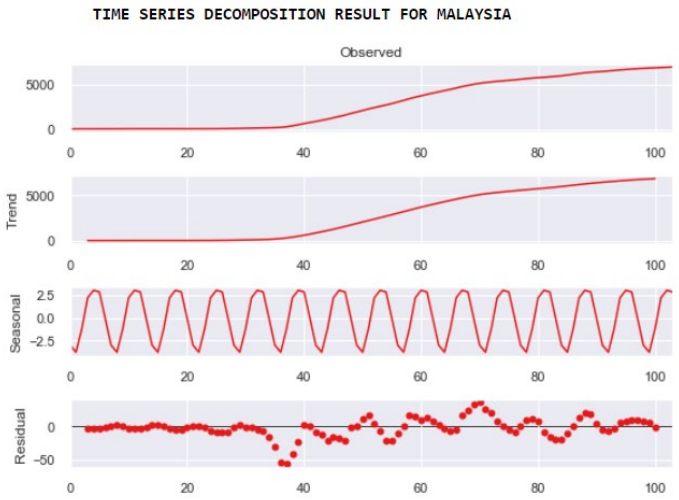
COVID-19 Datasets

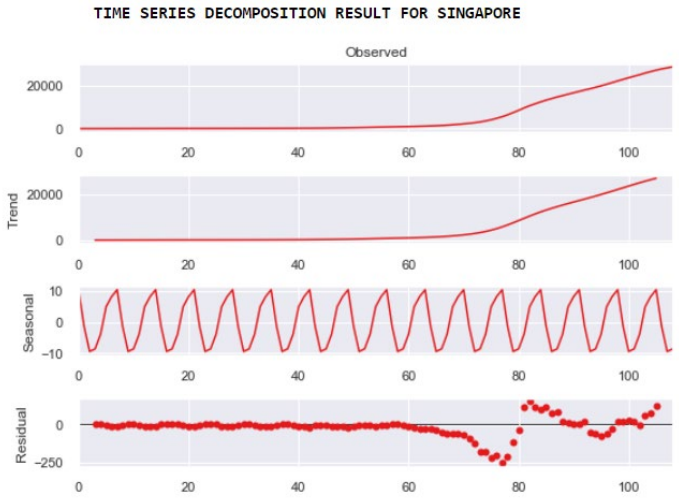
Interpreting and Communicating Data Insights – Data Mining & Interpretation (Descriptive Analytics)

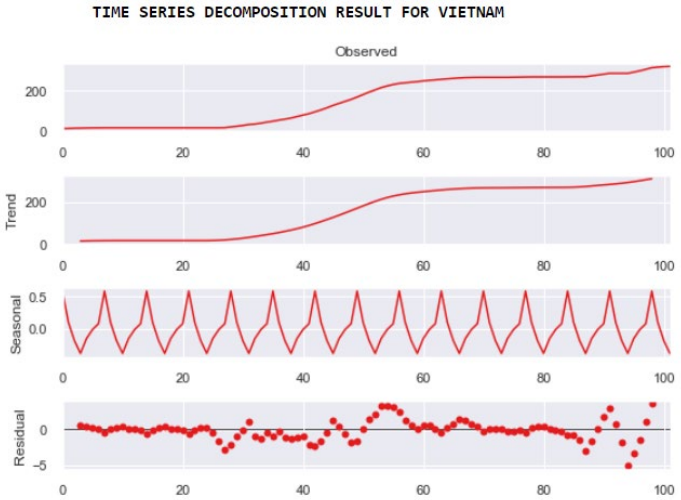
UNIVERSITY OF MALAYA

COVID-19 Datasets

Interpreting and Communicating Data Insights – Time Series Decomposition Evaluation (Diagnostic Analytics)







MALAYSIA

Results of Dickey-Fuller Test:	
Test Statistic	-0.887220
p-value	0.792152
#Lags Used	13.000000
Number of Observations Used	90.000000
Critical Value (1%)	-3.505190
Critical Value (5%)	-2.894232
Critical Value (10%)	-2.584210
dtype: float64	

SINGAPORE

Results of Dickey-Fuller Test:	
Test Statistic	-1.374493
p-value	0.594442
#Lags Used	13.000000
Number of Observations Used	95.000000
Critical Value (1%)	-3.501137
Critical Value (5%)	-2.892480
Critical Value (10%)	-2.583275
dtype: float64	

VIETNAM

Results of Dickey-Fuller Test:	
Test Statistic	-0.864807
p-value	0.799341
#Lags Used	13.000000
Number of Observations Used	88.000000
Critical Value (1%)	-3.506944
Critical Value (5%)	-2.894990
Critical Value (10%)	-2.584615
dtype: float64	

COVID-19 Datasets

Interpreting and Communicating Data Insights – Time Series Decomposition Evaluation (Diagnostic Analytics)

CHINA

Results of Dickey-Fuller Test:

Test Statistic	-4.415207
p-value	0.000279
#Lags Used	13.000000
Number of Observations Used	103.000000
Critical Value (1%)	-3.495493
Critical Value (5%)	-2.890037
Critical Value (10%)	-2.581971

dtype: float64

US

Results of Dickey-Fuller Test:

Test Statistic	0.186853
p-value	0.971514
#Lags Used	9.000000
Number of Observations Used	95.000000
Critical Value (1%)	-3.501137
Critical Value (5%)	-2.892480
Critical Value (10%)	-2.583275

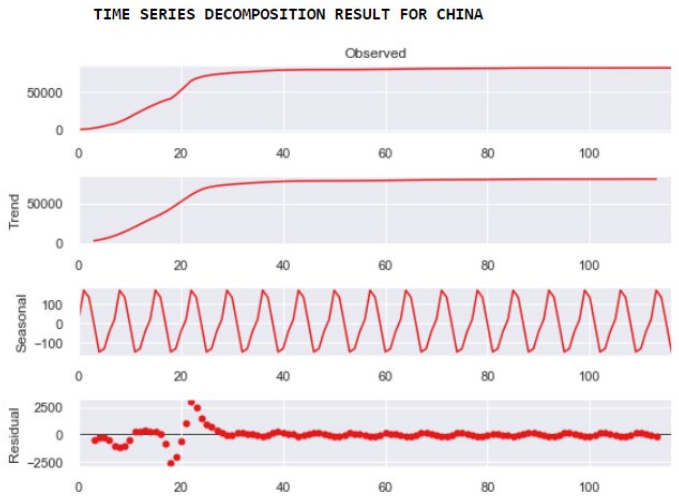
dtype: float64

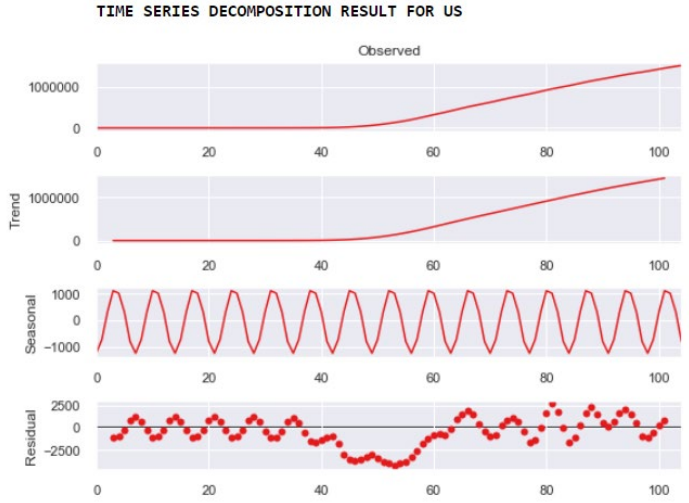
UK

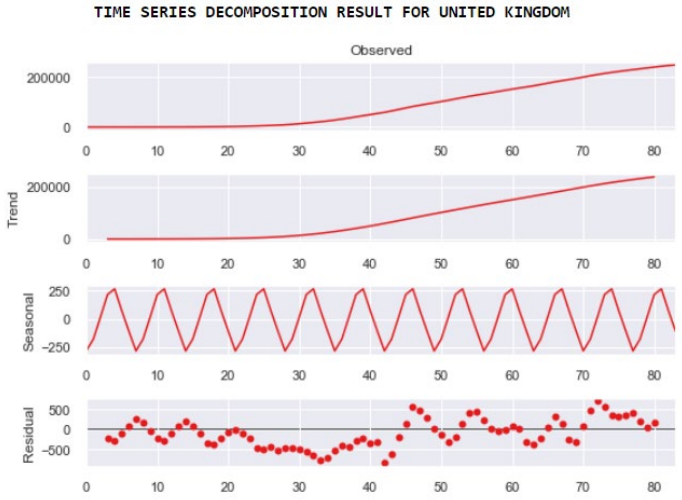
Results of Dickey-Fuller Test:

Test Statistic	-1.009380
p-value	0.749828
#Lags Used	10.000000
Number of Observations Used	73.000000
Critical Value (1%)	-3.523284
Critical Value (5%)	-2.902031
Critical Value (10%)	-2.588371

dtype: float64



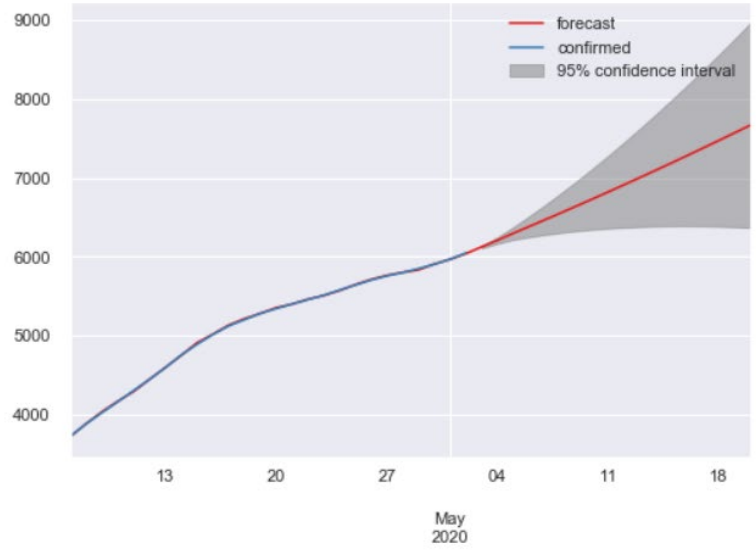


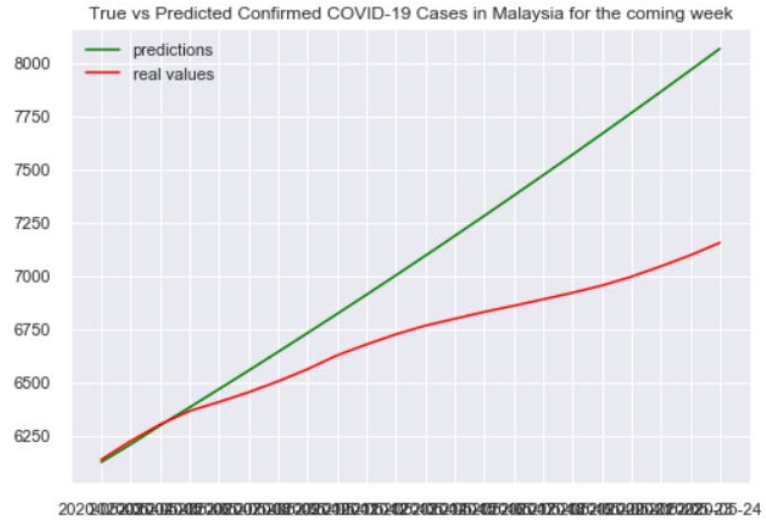


COVID-19 Datasets

Interpreting and Communicating Data Insights – Confirmed Cases Prediction for Malaysia (Predictive Analytics)

Mean absolute percentage error: 9.234862





ARIMA Model Results

Dep. Variable:	D2.confirmed	No. Observations:	84
Model:	ARIMA(0, 2, 4)	Log Likelihood	-298.514
Method:	css-mle	S.D. of innovations	7.985
Date:	Mon, 25 May 2020	AIC	609.028
Time:	14:29:21	BIC	623.613
Sample:	02-09-2020 - 05-02-2020	HQIC	614.891

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.9184	1.619	0.567	0.570	-2.254	4.091
ma.L1.D2.confirmed	0.4501	0.094	4.770	0.000	0.265	0.635
ma.L2.D2.confirmed	0.4614	0.105	4.395	0.000	0.256	0.667
ma.L3.D2.confirmed	0.4827	0.103	4.672	0.000	0.280	0.685
ma.L4.D2.confirmed	-0.5286	0.094	-5.624	0.000	-0.713	-0.344

Roots

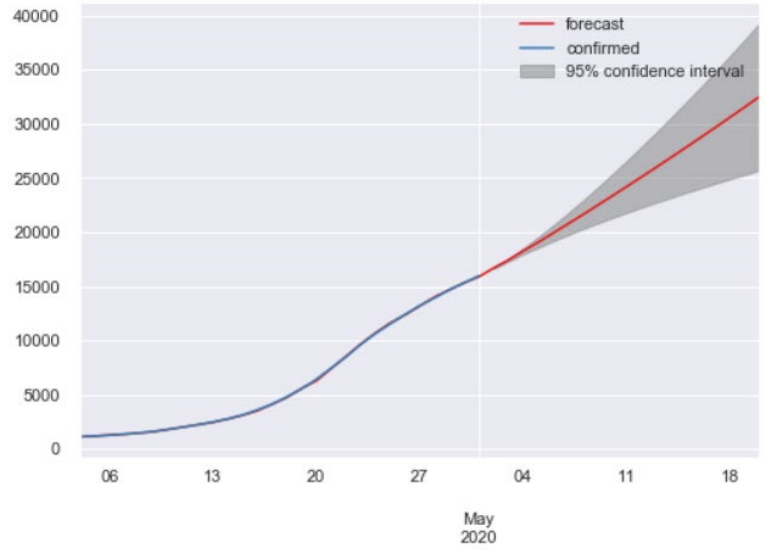
	Real	Imaginary	Modulus	Frequency
MA.1	-1.0000	-0.0000j	1.0000	-0.5000
MA.2	0.0106	-0.9999j	1.0000	-0.2483
MA.3	0.0106	+0.9999j	1.0000	0.2483
MA.4	1.8919	-0.0000j	1.8919	-0.0000

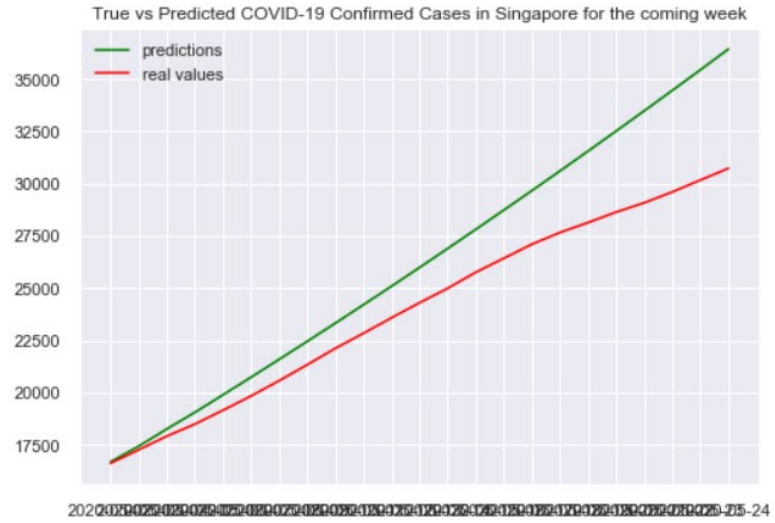
Mean absolute percentage error: 9.234862

COVID-19 Datasets

Interpreting and Communicating Data Insights – Confirmed Cases Prediction for Singapore (Predictive Analytics)

```
Mean absolute percentage error: 28.230667
```





ARIMA Model Results

Dep. Variable:	D2.confirmed	No. Observations:	88
Model:	ARIMA(0, 2, 5)	Log Likelihood	-424.134
Method:	css-mle	S.D. of innovations	28.283
Date:	Mon, 25 May 2020	AIC	862.267
Time:	14:30:04	BIC	879.609
Sample:	02-04-2020 - 05-01-2020	HQIC	869.254

	coef	std err	z	P>|z|	[0.025	0.975]
const	8.8241	8.296	1.064	0.287	-7.435	25.084
ma.L1.D2.confirmed	0.3547	0.127	2.796	0.005	0.106	0.603
ma.L2.D2.confirmed	0.6981	0.108	6.481	0.000	0.487	0.909
ma.L3.D2.confirmed	0.7165	0.109	6.584	0.000	0.503	0.930
ma.L4.D2.confirmed	-0.3004	0.099	-3.046	0.002	-0.494	-0.107
ma.L5.D2.confirmed	0.3266	0.110	2.969	0.003	0.111	0.542

Roots

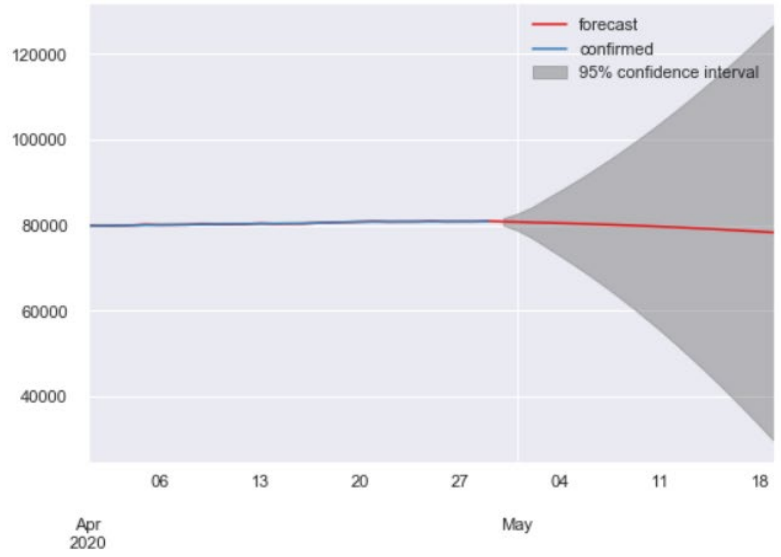
	Real	Imaginary	Modulus	Frequency
MA.1	-1.0000	-0.0000j	1.0000	-0.5000
MA.2	0.0137	-0.9999j	1.0000	-0.2478
MA.3	0.0137	+0.9999j	1.0000	0.2478
MA.4	0.9462	-1.4720j	1.7499	-0.1591
MA.5	0.9462	+1.4720j	1.7499	0.1591

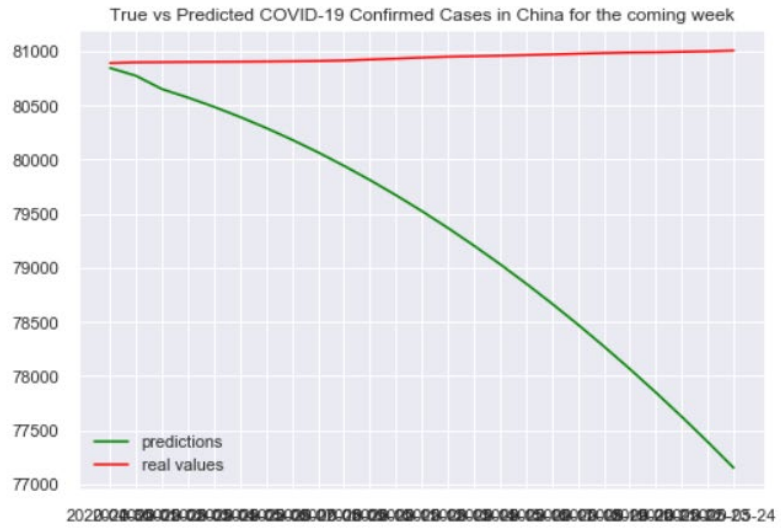
```
Mean absolute percentage error: 28.230667
```

COVID-19 Datasets

Interpreting and Communicating Data Insights – Confirmed Cases Prediction for China (Predictive Analytics)

Mean absolute percentage error: 1.996817





ARIMA Model Results

Dep. Variable:	D2.confirmed	No. Observations:	94
Model:	ARIMA(0, 2, 4)	Log Likelihood	-699.312
Method:	css-mle	S.D. of innovations	397.707
Date:	Mon, 25 May 2020	AIC	1410.625
Time:	14:31:29	BIC	1425.884
Sample:	01-27-2020 - 04-29-2020	HQIC	1416.789

	coef	std err	z	P>|z|	[0.025	0.975]
const	-7.7278	44.929	-0.172	0.863	-95.787	80.331
ma.L1.D2.confirmed	0.2940	0.092	3.192	0.001	0.113	0.474
ma.L2.D2.confirmed	0.2247	0.096	2.345	0.019	0.037	0.412
ma.L3.D2.confirmed	0.2458	0.094	2.627	0.009	0.062	0.429
ma.L4.D2.confirmed	-0.6849	0.092	-7.451	0.000	-0.865	-0.505

Roots

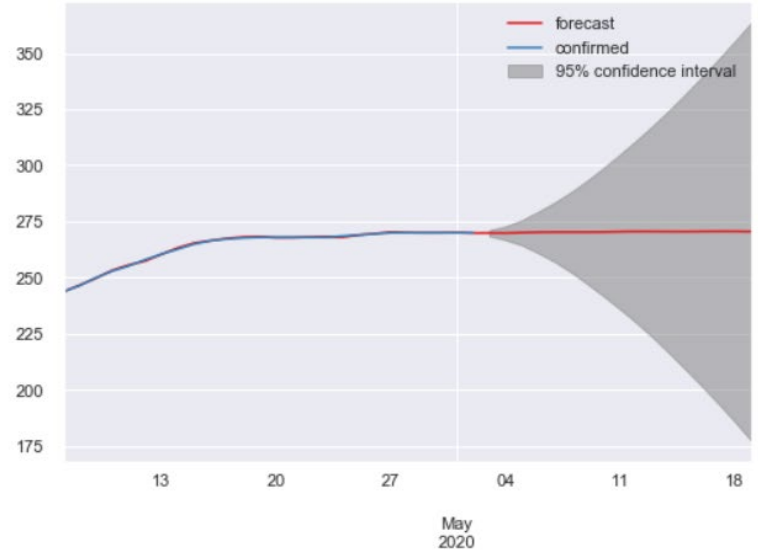
	Real	Imaginary	Modulus	Frequency
MA.1	-1.0000	-0.0000j	1.0000	-0.5000
MA.2	-0.0101	-1.0289j	1.0289	-0.2516
MA.3	-0.0101	+1.0289j	1.0289	0.2516
MA.4	1.3791	-0.0000j	1.3791	-0.0000

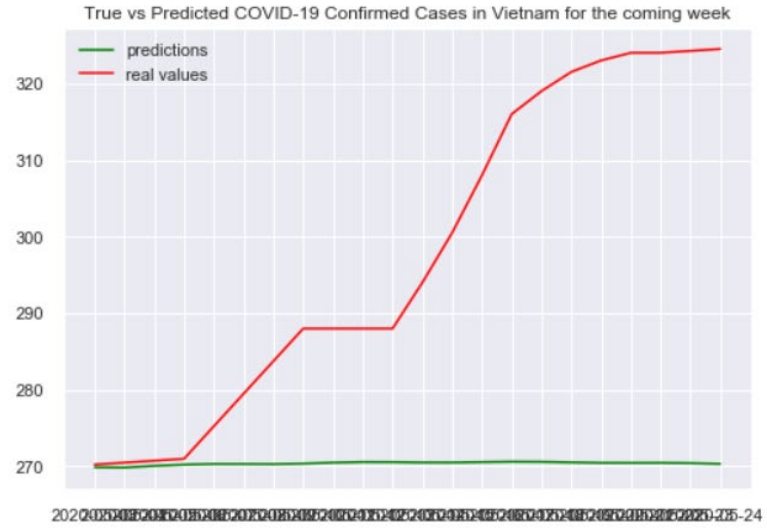
Mean absolute percentage error: 1.996817

COVID-19 Datasets

Interpreting and Communicating Data Insights – Confirmed Cases Prediction for Vietnam (Predictive Analytics)

Mean absolute percentage error: 8.781584





ARIMA Model Results

Dep. Variable:	D2.confirmed	No. Observations:	82
Model:	ARIMA(4, 2, 5)	Log Likelihood	-88.356
Method:	css-mle	S.D. of innovations	0.662
Date:	Mon, 25 May 2020	AIC	198.711
Time:	14:32:49	BIC	225.185
Sample:	02-11-2020 - 05-02-2020	HQIC	209.340

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0078	0.134	-0.059	0.953	-0.270	0.255
ar.L1.D2.confirmed	0.1557	0.130	1.201	0.230	-0.098	0.410
ar.L2.D2.confirmed	-0.6959	0.113	-6.167	0.000	-0.917	-0.475
ar.L3.D2.confirmed	-0.3717	0.108	-3.446	0.001	-0.583	-0.160
ar.L4.D2.confirmed	-0.2033	0.130	-1.561	0.119	-0.459	0.052
ma.L1.D2.confirmed	-0.0117	0.089	-0.131	0.896	-0.186	0.163
ma.L2.D2.confirmed	1.0088	0.117	8.602	0.000	0.779	1.239
ma.L3.D2.confirmed	1.0121	0.106	9.565	0.000	0.805	1.219
ma.L4.D2.confirmed	-0.0550	0.101	-0.543	0.587	-0.253	0.143
ma.L5.D2.confirmed	0.9533	0.122	7.823	0.000	0.714	1.192

Roots

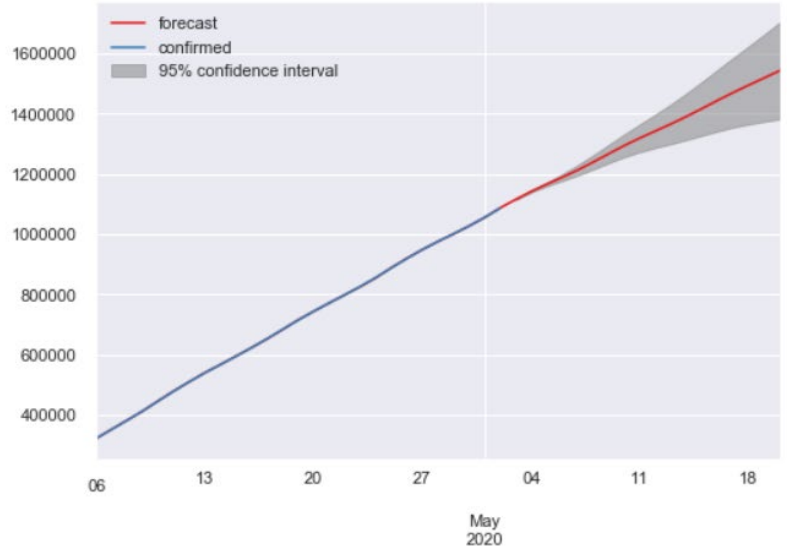
	Real	Imaginary	Modulus	Frequency
AR.1	0.4084	-0.9640j	1.0470	-0.1862
AR.2	0.4084	+0.9640j	1.0470	0.1862
AR.3	-1.3223	-1.6547j	2.1181	-0.3573
AR.4	-1.3223	+1.6547j	2.1181	0.3573
MA.1	-1.0000	-0.0000j	1.0000	-0.5000
MA.2	0.4932	-0.8976j	1.0242	-0.1700
MA.3	0.4932	+0.8976j	1.0242	0.1700
MA.4	0.0357	-0.9994j	1.0000	-0.2443
MA.5	0.0357	+0.9994j	1.0000	0.2443

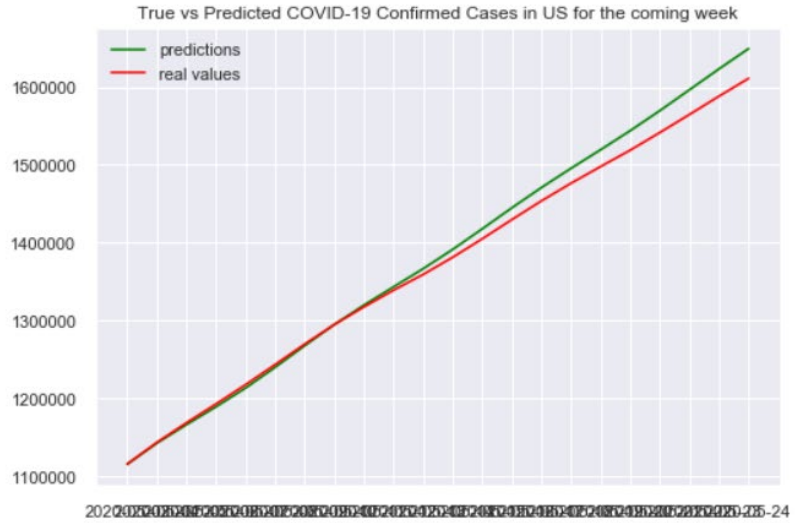
Mean absolute percentage error: 8.781584

COVID-19 Datasets

Interpreting and Communicating Data Insights – Confirmed Cases Prediction for US (Predictive Analytics)

Mean absolute percentage error: 13.295892





ARIMA Model Results

Dep. Variable:	D2.confirmed	No. Observations:	85
Model:	ARIMA(4, 2, 4)	Log Likelihood	-649.430
Method:	css-mle	S.D. of innovations	467.235
Date:	Mon, 25 May 2020	AIC	1318.860
Time:	14:36:02	BIC	1343.287
Sample:	02-08-2020	HQIC	1328.685
	- 05-02-2020		

	coef	std err	z	P>|z|	[0.025	0.975]
const	212.3919	331.488	0.641	0.522	-437.313	862.097
ar.L1.D2.confirmed	1.2543	0.113	11.142	0.000	1.034	1.475
ar.L2.D2.confirmed	-0.3253	0.177	-1.838	0.066	-0.672	0.022
ar.L3.D2.confirmed	-0.6906	0.174	-3.971	0.000	-1.031	-0.350
ar.L4.D2.confirmed	0.6597	0.099	6.690	0.000	0.466	0.853
ma.L1.D2.confirmed	-0.2358	0.095	-2.483	0.013	-0.422	-0.050
ma.L2.D2.confirmed	0.1245	0.081	1.533	0.125	-0.035	0.284
ma.L3.D2.confirmed	0.5960	0.083	7.214	0.000	0.434	0.758
ma.L4.D2.confirmed	-0.7644	0.095	-8.032	0.000	-0.951	-0.578

Roots

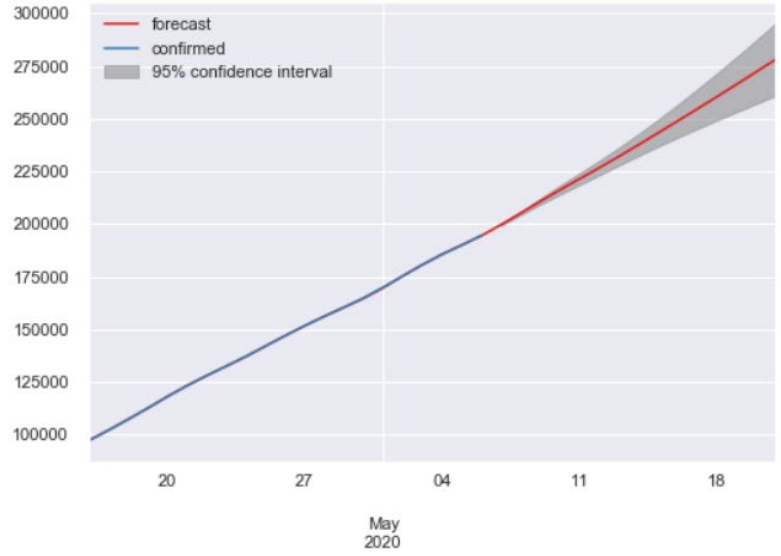
	Real	Imaginary	Modulus	Frequency
AR.1	-1.2681	-0.0000j	1.2681	-0.5000
AR.2	0.6184	-0.8523j	1.0530	-0.1501
AR.3	0.6184	+0.8523j	1.0530	0.1501
AR.4	1.0781	-0.0000j	1.0781	-0.0000
MA.1	-1.0000	-0.0000j	1.0000	-0.5000
MA.2	0.2357	-0.9718j	1.0000	-0.2121
MA.3	0.2357	+0.9718j	1.0000	0.2121
MA.4	1.3082	-0.0000j	1.3082	-0.0000

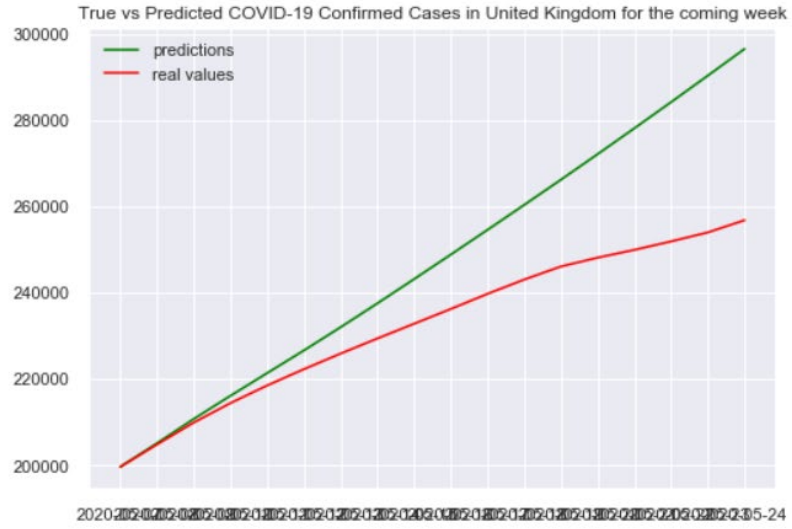
Mean absolute percentage error: 13.295892

COVID-19 Datasets

Interpreting and Communicating Data Insights – Confirmed Cases Prediction for UK (Predictive Analytics)

Mean absolute percentage error: 13.331572





ARIMA Model Results

Dep. Variable:	D2.confirmed	No. Observations:	68
Model:	ARIMA(4, 2, 5)	Log Likelihood	-453.216
Method:	css-mle	S.D. of innovations	171.416
Date:	Mon, 25 May 2020	AIC	928.431
Time:	14:36:38	BIC	952.846
Sample:	02-29-2020 - 05-06-2020	HQIC	938.105

	coef	std err	z	P>|z|	[0.025	0.975]
const	68.0934	49.839	1.366	0.172	-29.590	165.777
ar.L1.D2.confirmed	0.9459	0.134	7.071	0.000	0.684	1.208
ar.L2.D2.confirmed	-0.0367	0.175	-0.210	0.834	-0.380	0.306
ar.L3.D2.confirmed	-0.4039	0.168	-2.408	0.016	-0.733	-0.075
ar.L4.D2.confirmed	0.2604	0.135	1.930	0.054	-0.004	0.525
ma.L1.D2.confirmed	-0.7963	0.104	-7.642	0.000	-1.001	-0.592
ma.L2.D2.confirmed	0.0972	0.120	0.810	0.418	-0.138	0.332
ma.L3.D2.confirmed	0.0972	0.121	0.805	0.421	-0.139	0.334
ma.L4.D2.confirmed	-0.7963	0.122	-6.526	0.000	-1.035	-0.557
ma.L5.D2.confirmed	1.0000	0.095	10.497	0.000	0.813	1.187

Roots

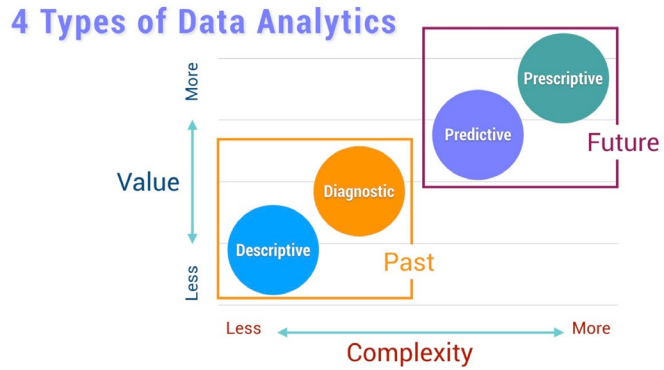
	Real	Imaginary	Modulus	Frequency
AR.1	-1.4640	-0.0000j	1.4640	-0.5000
AR.2	0.8688	-1.1394j	1.4329	-0.1463
AR.3	0.8688	+1.1394j	1.4329	0.1463
AR.4	1.2778	-0.0000j	1.2778	-0.0000
MA.1	-1.0000	-0.0000j	1.0000	-0.5000
MA.2	-0.0287	-0.9996j	1.0000	-0.2546
MA.3	-0.0287	+0.9996j	1.0000	0.2546
MA.4	0.9269	-0.3753j	1.0000	-0.0612
MA.5	0.9269	+0.3753j	1.0000	0.0612

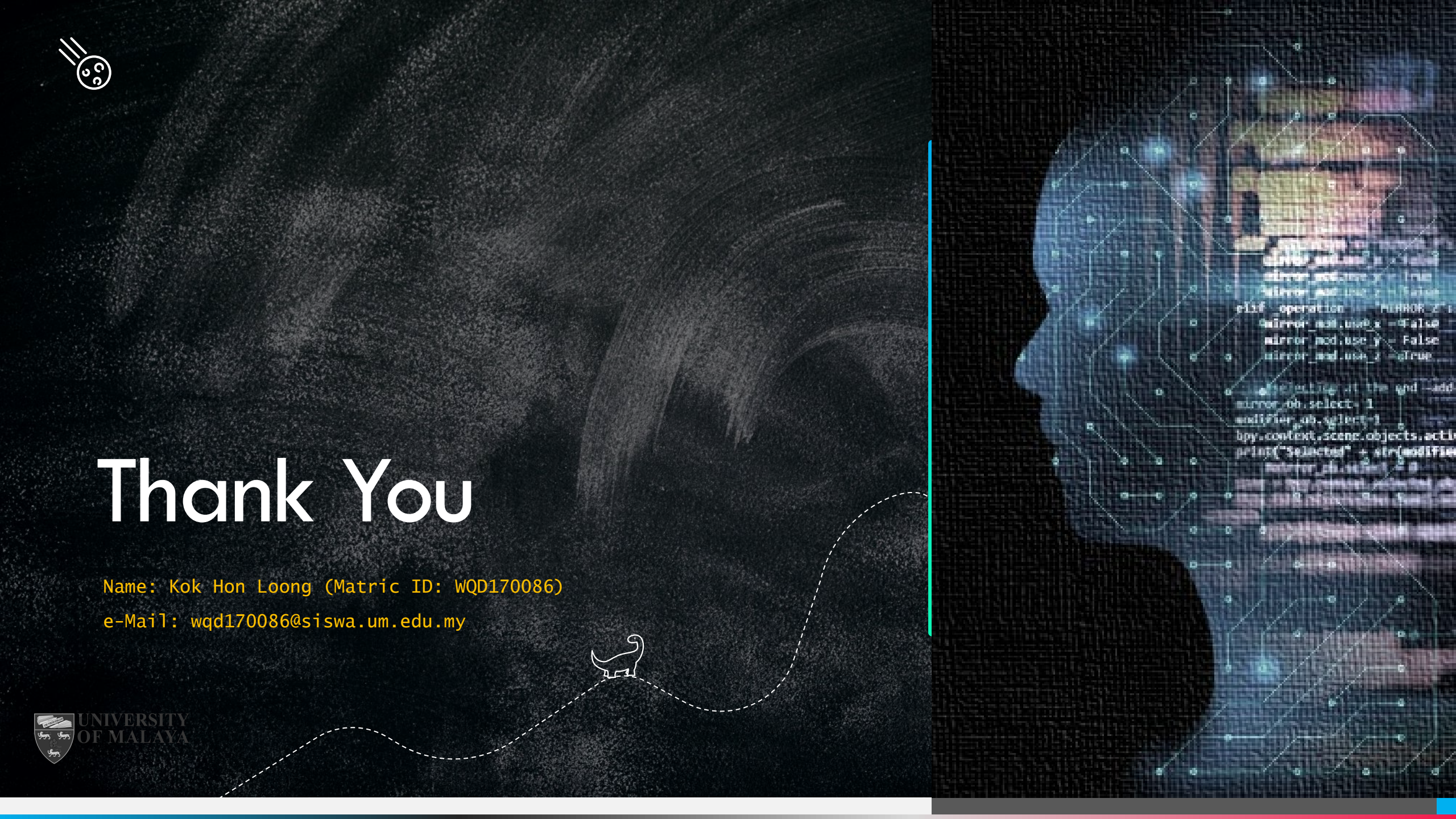
Mean absolute percentage error: 13.331572

Summary

- Data analysis and interpretation have now **taken center stage** with the advent of the digital age.
- Capable of displaying key performance indicators (KPIs) for both **quantitative and qualitative data analyses**, are ideal for making the fast-paced and data-driven market decisions with sustainable success.
- Data interpretation refers to the implementation of processes through which **data is reviewed** for the purpose of arriving at an **informed conclusion**.

- The interpretation of data **assigns a meaning to the information analyzed** and determines its signification and implications.
- Facts and figures are meaningless if you can't gain valuable insights that lead to more-informed actions. Analytics solutions offer a convenient way to leverage business data**:**
  - ☒ **Descriptive**: tells what happened in the past
  - ☒ **Diagnostic**: helps understand why something happened in the past
  - ☒ **Predictive**: predicts what is most likely to happen in the future.
  - ☒ **Prescriptive**: recommends actions you can take to affect those outcomes.





Thank You

Name: Kok Hon Loong (Matric ID: WQD170086)

e-Mail: wqd170086@siswa.um.edu.my

UNIVERSITY OF MALAYA