ASSIGNMENT REPORT:

COVID-19 ANALYSIS AND CASES PREDICTION

NIK FAIZ AFIQ BIN NIK AB RAHMAN

WQD7006, Machine learning

UNIVERSITY OF MALAYA

Kuala lumpur

2020/2021 Semester

Table of Contents

[Table of Contents 2](#_Toc61877935)

[INTRODUCTION 3](#_Toc61877936)

[CHAPTER 1: Analysis and design 4](#_Toc61877937)

[1.1 Design 4](#_Toc61877938)

[1.2 Data Preprocessing /Preparation 4](#_Toc61877939)

[1.3 Exploratory Data Analysis (for secondary objective) 5](#_Toc61877940)

[1.4 Modeling and Evaluation 6](#_Toc61877941)

[CHAPTER 2: Experiment Result 8](#_Toc61877942)

[2.1 Input 8](#_Toc61877943)

[2.2 Output for objective 2 9](#_Toc61877944)

[2.3 Output for objective 1 12](#_Toc61877945)

[CHAPTER 3: Discussion (Strength and limitation) 15](#_Toc61877946)

[3.1 Analysis on objective 2 output 15](#_Toc61877947)

[3.2 Analysis on objective 1 output 15](#_Toc61877948)

[CHAPTER 4: Conclusion 16](#_Toc61877949)

[References 17](#_Toc61877950)

INTRODUCTION

The outbreak of coronavirus disease (COVID-19) has been declared a Public Health Emergency of International Concern (PHEIC) and the virus has now spread to many countries and territories. Based on recent data, it has affected more than 90 million people around the world and causing close to 2 million death.

COVID-19 is a disease caused by a new strain of coronavirus. The COVID-19 virus is a new virus linked to the same family of viruses as Severe Acute Respiratory Syndrome (SARS) and some types of common cold. Symptoms can include fever, cough and shortness of breath. In more severe cases, infection can cause pneumonia or breathing difficulties. More rarely, the disease can be fatal. The virus is transmitted through direct contact with respiratory droplets of an infected person (generated through coughing and sneezing).

The main objective of this project is to predict the number of new COVID-19 Cases based on previous history of cases. This will enable us to see the current trend of this cases and take an appropriate action towards it. The secondary objective of this project is to classify the coronavirus trend for each of the countries based on the countries’ new cases in last 7 days. From this, we can see which of the countries are handling covid-19 effectively and which country is doing really bad. After all, the first step to any improvement is realizing the current situation of COVID-19 in every country.

For this, we are using the data provided by John Hopkins University that is publicly available to the people. People can access it through this given link: <https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data>.

# Analysis and design

## Design

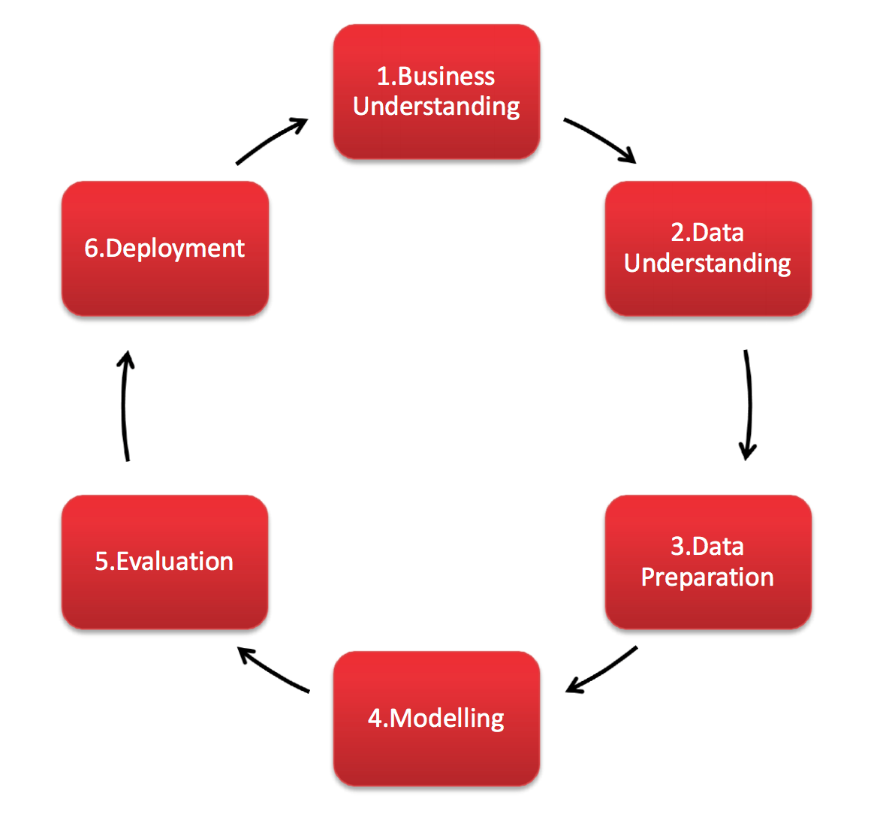


Figure 1‑1: CRISP-DM Methodology

This assignment report is designed to follow the CRISP-DM methodology that provides a structured approach to planning a data mining project. In business understanding, we try to understand the issues and provide an objective based on that issue. Then from there, we search for a suitable dataset to tackle this issue and achieve the objective (data understanding). The data is then prepared and undergoing pre-processing process (data preparation). Then we model the data using machine learning algorithm (modeling). We will then evaluate the models accordingly (evaluation). However, this model is yet to be deployed to be used in real life yet. All the codes that is used in this assignment report is available in this GitHub link: <https://github.com/n-Faiz/WQD7006_ML_Covid19>.

## Data Preprocessing /Preparation

The data input consists of 3 csv files, each contains data of COVID-19 number of confirmed cases, number of death and number of people recovered respectively. Each of the data have list of countries as the rows with the dates as the column. Since we need the data to be in a nice tabular format with column as the attributes, we will need to do some pre-processing on the data. First, we will need to unpivot the date columns so all the numbers for that attribute is under the same column. Unpivoting here will be done on all there of the dataset. Example of unpivoting: for the numbers of confirmed cases, instead of the columns originally have different dates as column, we will now convert it so that the dates now occupy only 1 of the columns and the other columns fills the value corresponding to the dates and the countries.

After that, all 3 tables now need to be merged together for easier analysis. The data is then checked for any null values and an appropriate action are taken for these null values. Since the dataset only consist of total new confirmed cases, total death and total people recovered, the total active cases are then calculated by using the formula: total active = total confirmed – total death – total recovered. It is also useful to include daily new confirmed cases, daily new death and daily new recovered, so we include that in by taking the current total cases minus the cases on the previous day for that country.

In summary, for preprocessing, these steps are taken for data pre-processing:

1. Data Merging
2. Needs to be unpivoted (merge the dates to 1 column only)
3. Check for missing value
4. Calculate for new columns: Total Active
5. Calculate for new columns: New Confirmed, New Death, and New Recovered

## Exploratory Data Analysis (for secondary objective)

Our secondary objective of this project is to classify the coronavirus trend for each of the countries based on the countries’ new cases in last 7 days. For that, we need to use the numbers of new cases only and only include the data for the last seven days. Then we standardize it by dividing it with the maximum number of cases in the last 7 days for each country. This resulted the gradient produced on the next step is based on standardized unit since the range of y axis is the same that is from 0 to 1 for every country. Next the gradient is calculated based on the 7 data point for each country. Then, categorized the trend of the cases in each country based on the gradient. Here we defined that the trend is “increasing” if the gradient value is more than 0.2, “around the same” if the gradient value is between -0.2 and 0.2, “decreasing” if the value is below -0.2 and “No Cases” if there is no new cases observed for the last 7 days. Then the number of countries with each status is visualized using bar chart. Then we will plot the performance of the top 10 of “decreasing” trend to see the graph of the top 10 countries that performs best to combat covid-19 within the last 7 days and the same for “increasing” trend.

Summary of Exploratory Data Analysis steps taken:

1. Filter the data to only include last 7 days
2. Standardize it based on the country before calculating the gradient (the new confirmed cases are standardized so it falls within 0 and 1 in each country)
3. Calculate the gradient for each of the country based on their last 7 new cases
4. Categorize the status based on the gradient
5. Visualize total country that have increasing, decreasing or same gradient
6. Plot top 10 performers and bottom 10 performers

## Modeling and Evaluation

The models that we used here for prediction and forecasting are: Support Vector Machine (SVM), Simple Linear Regression, Polynomial Linear Regression and Decision Tree regression. But before the modeling is done, we need to split the data into train and test split. Here we are using 70% of the data as training while the remaining 30% of data as testing. We also provide 20 extra points that is used to forecasting 20 days into the future. Instead of using the date as the X value, integer values are used to make it easy for calculation. The model is then evaluated using 3 score evaluation: Mean Absolute Value, Root Mean Square Error, and R2 score. The result is also projected into a graph so we can see the points clearly.

# Experiment Result

## Input

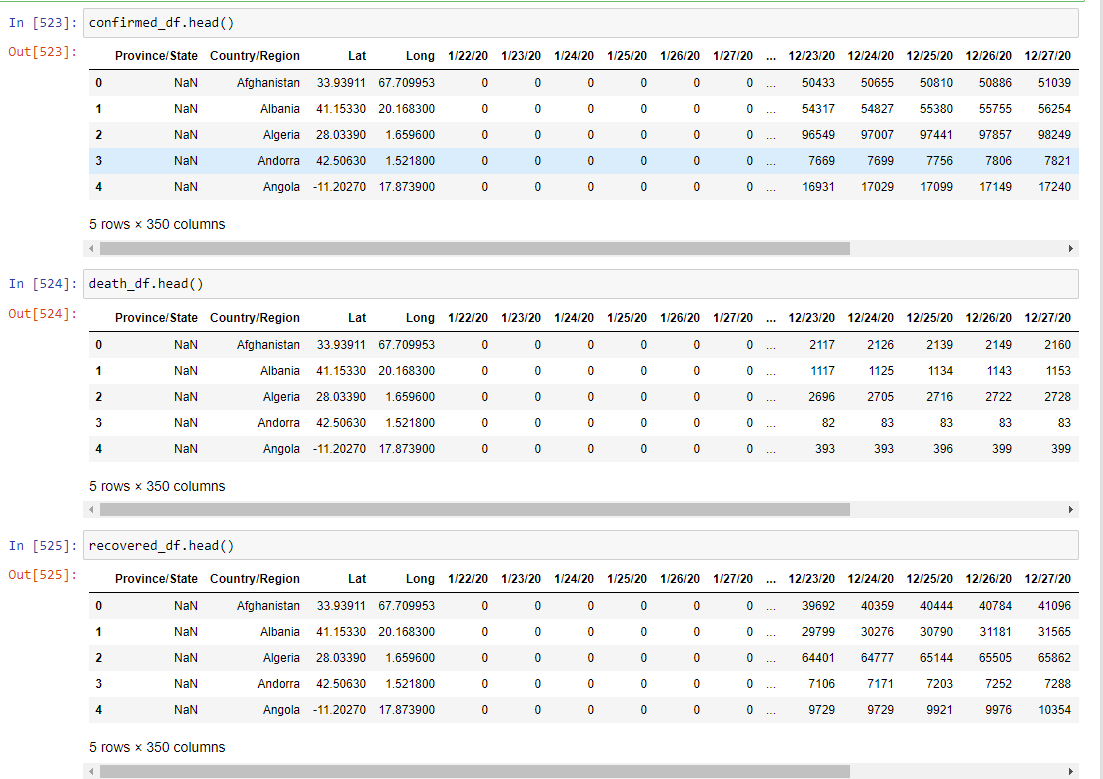


Figure 2‑1: Original Data

Our source files are originally the 3 comma separated value format file. The first one contains the confirmed cases of covid-19, the second file contains the number of deaths due to COVID-19 in each country and the last one contains the number of recovered people from COVID-19. Each of the files have the countries and regions as the rows and the dates as the column.

## Output for objective 2

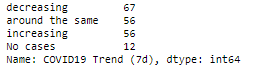


Figure 2‑2: COVID-19 trend in each country

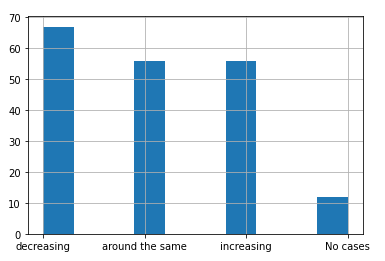


Figure 2‑3: COVID-19 Trend in each country

As of 18th January 2021, we can see that most of the countries have decreasing trend of new COVID-19 cases. There are 67 countries that have decreasing trend, 56 countries have flat trend (around the same), 67 countries that have increasing COVID-19 new cases trend, and 12 countries that have no new cases at all.

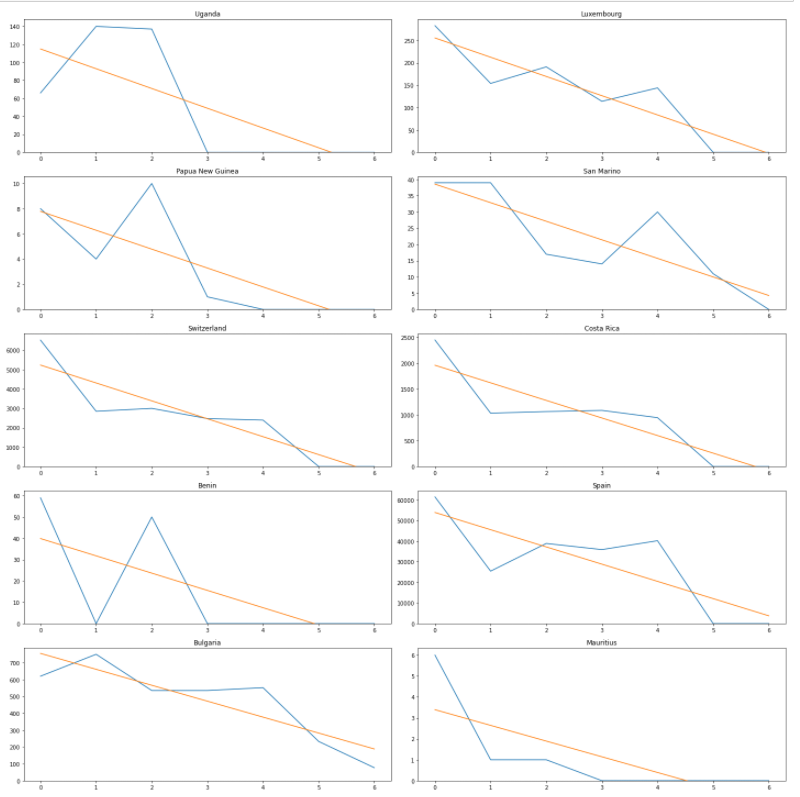


Figure 2‑4: Top 10 countries with decreasing trend of New COVID-19 cases

List of top 10 countries with decreasing trend of new COVID-19 cases: Uganda, Luxembourg, Papua New Guinea, San Marino, Switzerland, Costa Rica, Benin, Spain, Bulgaria and Mauritius. Some of them have 0 new cases for the last couple of days, a sign that these countries are combating COVID-19 effectively.

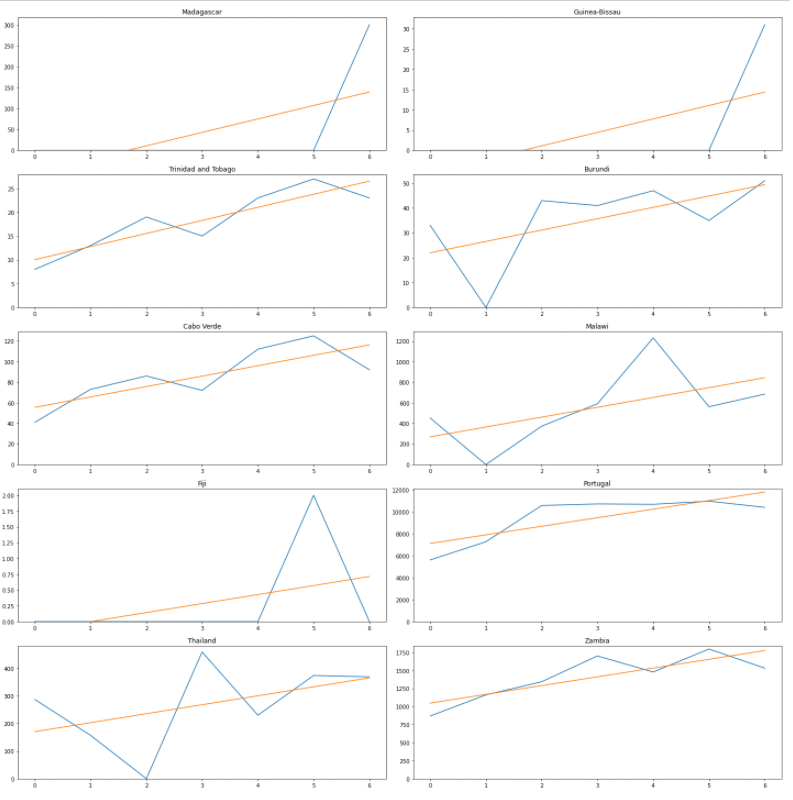


Figure 2‑5: Bottom 10 Countries with increasing trend of COVID-19 New Cases

These countries that have increasing trend for COVID-19 new cases in the last 7 days. These countries might need to take extra precautions to prevent further increase in COVID-19 cases.

## Output for objective 1

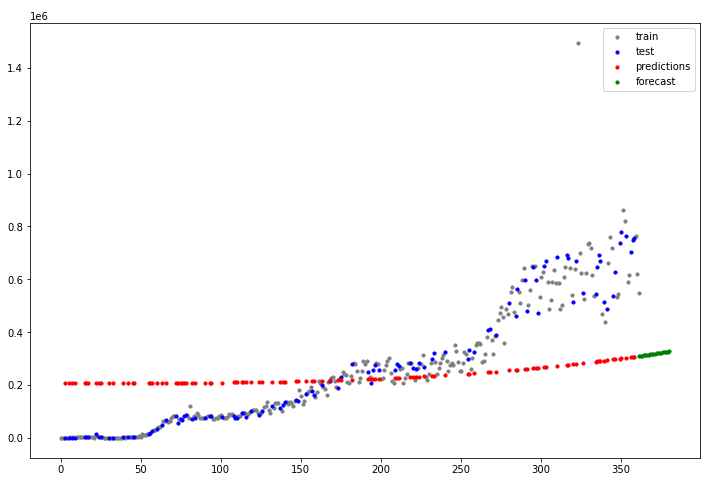


Figure 2‑6: SVR modeling and forecasting

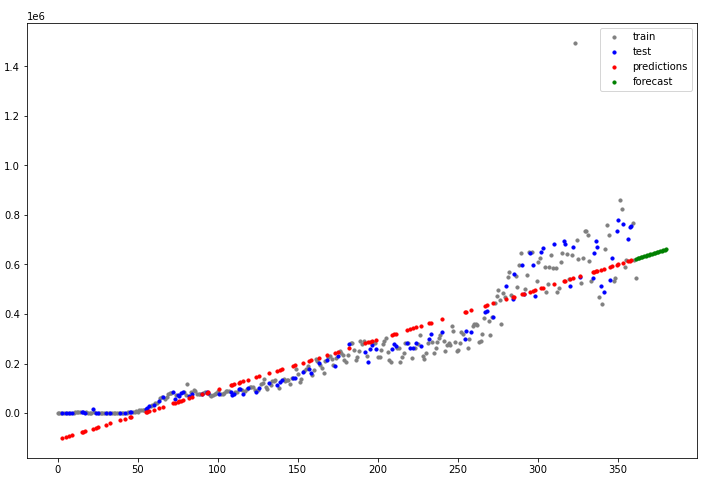


Figure 2‑8: Linear regression modeling and forecasting

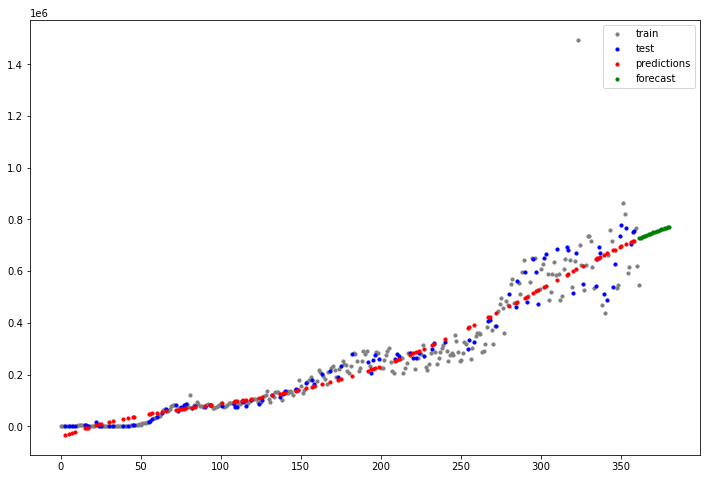


Figure 2‑9: Polynomial Regression modeling and forecasting

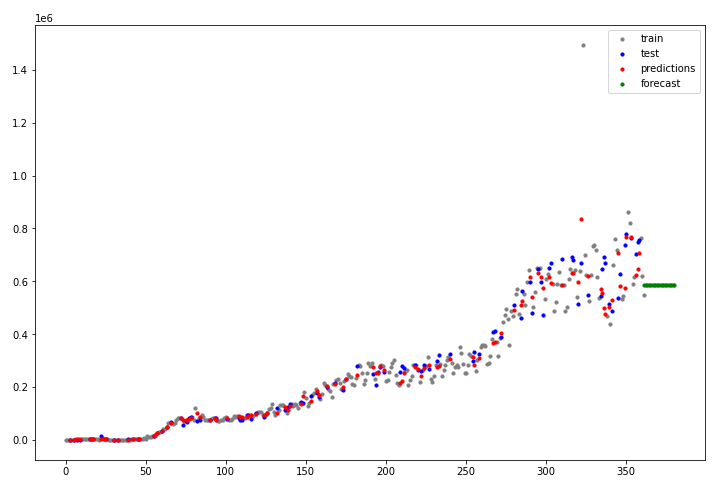


Figure 2‑10: Decision Tree Regressor Modeling and forecasting

Figure 2-7 to 2-10 shows the graph of the models that include the train value, the test value, the prediction value and the forecasting value, colored in grey, blue, red and green respectively. Strong estimator for prediction – regression tree and polynomial. SVR – fail. Linear ok. But for forecasting future value, decision tree is constant. Not good to forecasting future value.

Table 2‑1: Evaluation Score across different models

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Mean Absolute Error | Root Mean Square Error | R2 Score |
| Support Vector Machine | 177688.41 | 215668.40 | -44.62 |
| Simple Linear Regression | 58131.40 | 72334.89 | 0.90 |
| Polynomial Linear Regression | 34931.38 | 51009.30 | 0.95 |
| Random Forest | 27827.08 | 49588.08 | 0.96 |

Based on the evaluation score, Random Forest has the best score for all 3 of the evaluation score : lowest Mean Absolute Error, lowest Root Mean Square Error and highest R2 score. Here we can conclude that Random Forest is the best model for predicting the cases.

# Discussion (Strength and limitation)

performance of your method i.e. strength and limitations; expected results, why real outcome is different, suggestion for future improvements.

## Analysis on objective 2 output

Only limit to 7 days – volatile. Only study the new cases, exclude number of death, total cases, number of population, and recovery.

## Analysis on objective 1 output

Which model is the best for predicting test? Which model best for forecasting.

Some not good for forecasting. Some model even cant find a good fit for model. Further studies might be needed to address this. Can be in result?: SVR fail completely. D tree not good for forcasting but does great when predicting the value in between.

Can include more good model. Can use better parameter (currently using default parameter from scikit learn package.

Time series, might need to find better ways to test. Start with fitting with 10 data point, predict next 10 data point. Then fit 20 data point and predict next 10 data point and so on.

Can use other variable, total cases, total death, total recovered, new death, new recovered for predictions

# Conclusion

link practice to theory; image processing topics applied

lots of countries still have increasing new cases within the last 7 days. Worrying.

Can see the prediction total new cases worldwide is in increasing trend.

However, since the vaccine is already introduced and are used world-wide, hope this can reduce more.

References

Buchwalow, I. B., and Böcker, W. (2010). *Immunohistochemistry: basics and methods.* Berlin: Springer Verlag.

Caamaño-Tubío, R. I., Pérez, J., Ferreiro, S., and Aldegunde, M. (2007). Peripheral serotonin dynamics in the rainbow trout (*Oncorhynchus mykiss*). *Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology, 145*(2): 245-255

Cakir, Y., and Strauch, S. M. (2005). Tricaine (MS-222) is a safe anesthetic compound compared to benzocaine and pentobaritol to induce anesthesia in leopard frogs (*Rana pipiens*). *Pharmacological Reports, 57*: 467-474.

Cameron, A. A., Plenderleith, M. B. and Snow, P. J. (1990). Organization of the spinal cord in four species of elasmobranch fishes: cytoarchitecture and distribution of serotonin and selected neuropeptides. *The Journal of Comparative Neurology, 297*: 201-218