**Symbolic Computation Using Big Data Deliverables**

**Kayla Wright**

**CTEC 298**

**Dr. Bemley**

**Introduction**

The document covers the topics discussed in the CTEC 128 course for the project that covered the topic of world food production and the results from that research. The different materials submitted for this research includes a powerpoint, research questions, an excel spreadsheet containing the research data, and a final data science report. The data was shown through the use of plot visualization.

**Summary of 128 Paper**

The paper worked on in CTEC 128 covered World Food Production, and each team was given a scenario based on which topic they chose for their research. For World Food Production, the scenario was; Our world population is expected to grow from [7.3 billion today to 9.7 billion in the year 2050](http://www.un.org/en/development/desa/news/population/2015-report.html). Finding solutions for feeding the growing world population has become a hot topic for food and agriculture organizations, entrepreneurs, activists, and [philanthropists](https://canwefeedtheworld.wordpress.com/tag/bill-gates/). These solutions range from changing the way we [grow our food](https://www.forbes.com/sites/christinatroitino/2017/08/24/memphis-meats-lab-grown-meat-raises-17m-with-help-from-bill-gates-and-richard-branson/#2f8186d43fd0) to changing the [way we eat](https://www.peta.org/issues/animals-used-for-food/global-warming/). To make things harder, the world's climate is changing and it is both affecting and affected by the way we grow our food. You work for Feed the Hungry (FH), a non-profit organization dedicated to fighting hunger and poverty across the globe. FH is concerned about the availability of food for the world’s most vulnerable population. The organization’s leaders want to create a strategic plan for the future to ensure that there is enough food to feed everyone. As the organization’s data scientist team, your job is to give these leaders initial insights on worldwide food production and consumption.

After given the scenario, each team was given a data set to work with in order to narrow down some of their results for the research portion of the project. well. In the unwrangled data set given for world food production, there were a total number of 64 variables and a total number of 12,748 observations. The areas we chose to focus on were the United States, India, China, and Brazil because they were the top food producing areas based off of the dataset provided. The next thing was to put together a set of research questions which were, Why is there going to be such a population boost in 31 years? What do the world’s highest food countries produce and how? How much space do we need to farm and distribute the food? Why is it that some of the largest countries produce very little food? What kinds of food being produced are the easiest to grow? These questions help guide the research for how data should be gathered, and how to begin the research.

An important factor to know if what food each region produces the most of. Since corn was one of the top produced foods in the top food producing areas, it is important to find reasons why other regions should grow this crop to help world food production. Corn can be found in: starch, oil, food sweeteners, alcohol, as well as livestock feed and biofuel that assists global food security.

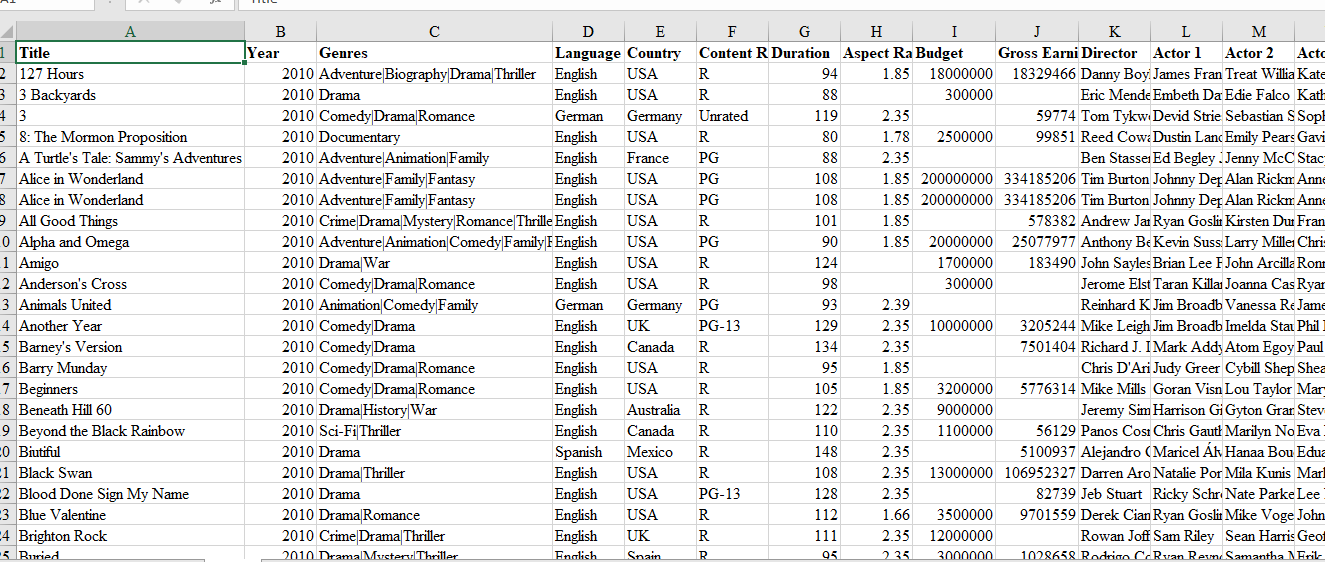
**Description of CTEC 128 Material Submitted**

The materials there were submitted for CTEC 128 were a PowerPoint, an excel spreadsheet with the data used for research for world food production, and a final data science paper. In the un-wrangled excel spreadsheet data that was submitted, there were a total number of 64 variables and a total number of 12,748 observations. The excel spreadsheet rows are labeled by the area abbreviation which is a shorter term for the area code, areas which are the different types of countries. Item code is a unique identifier assigned to each finished/manufactured product which is ready to be marketed or for sale. The item, the element code, the element which is either food or feed. Food is (uncountable) any substance that is or can be consumed by living organisms, especially by eating, in order to sustain life while feed is (uncountable) food given to (especially herbivorous) animals. The unit amount of food, latitude and longitude. And lastly, each of the years of food production. The powerpoint gives a visual of the graphs and excel spreadsheet. The final data science report submitted contains the overall research for the topic of world food production, and why the group chose the 4 different countries over all of the other countries.

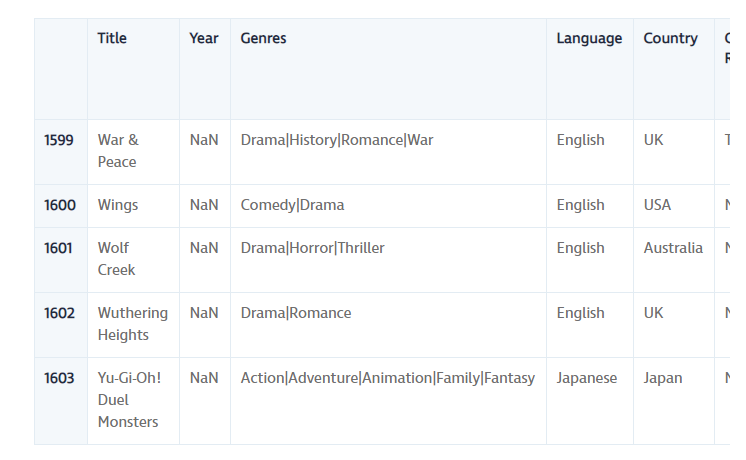
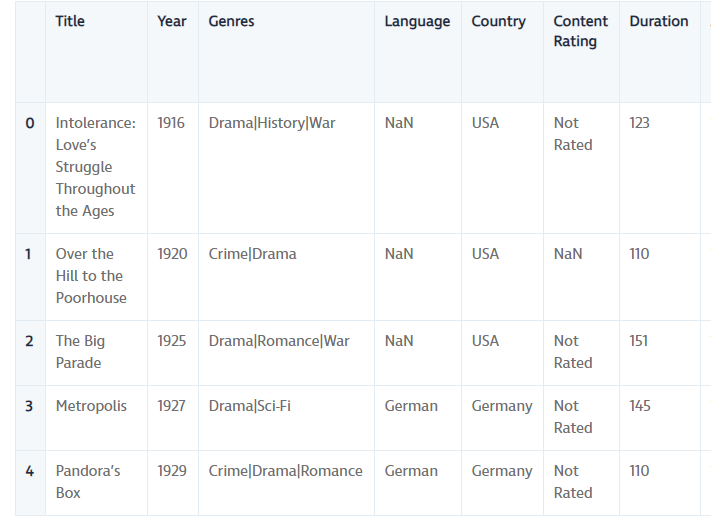
**Description of Plot Deliverables**

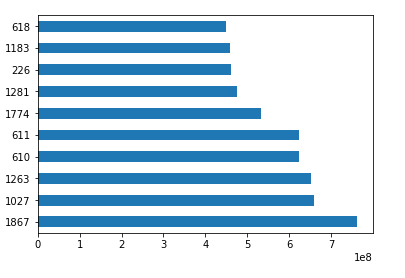
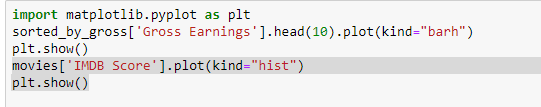
The use of data visualization gives you the ability to present data in a pictorial or graphical format. This allows you to analyze, and transform your data. The 6 different plots used for the plot deliverables consisted of, scatter plots, bar graphs, histograms, pie plot/chart, stack/area plot, and multiplot graph. Your bar graphs are used to compare data between groups and measure changes over time. A histogram is used to represent data given in the form of some groups. A scatter plot is used to compare two or more variables, or compare the changes over time from two groups. A pie chart is used to visualize categorical data. A stack plot is used to track changes over time. A multiplot graph shows subplots for 2 or more sets of data. In the documentation provided, there is an example and how to create each plot.

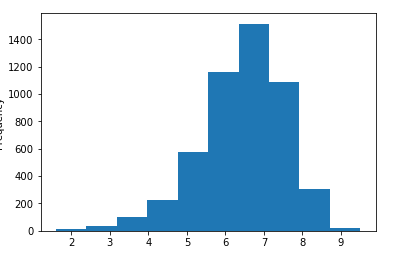
The original data set that I used was based off of “food production” but the csv file wouldn’t properly upload to the jupyter notebook program, so for this demonstration, I used a different csv files based off of movies.

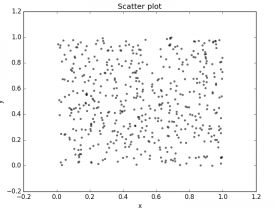


Before you are able to use any of your data, you have to upload the csv file to the jupyter notebook. In the first example of the bar chart, the movies.head() program will display the first few lines of the data. Head in the previous section that displays few rows from the top from the DataFrame. We can use the tail method to view the bottom rows. If no parameter is passed, only the bottom five rows are returned.

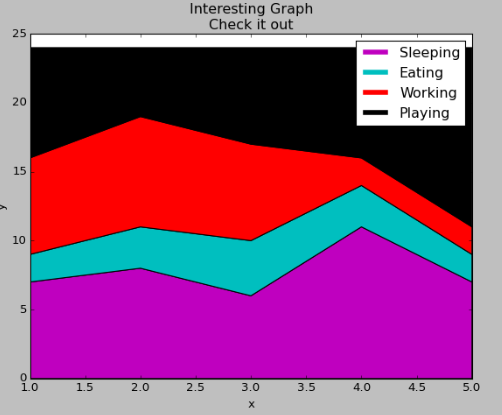
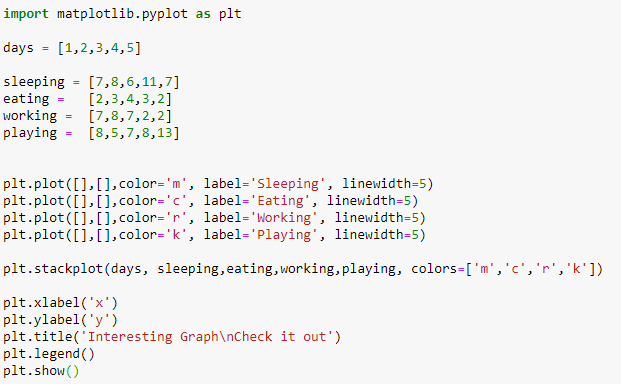


To create a histogram of the IMDB Scores to check the distribution of IMDB Scores across all movies. Histograms are a good way to visualize the distribution of a data set. We use the plot method on the IMDB Scores series from our movies DataFrame and pass it the argument.

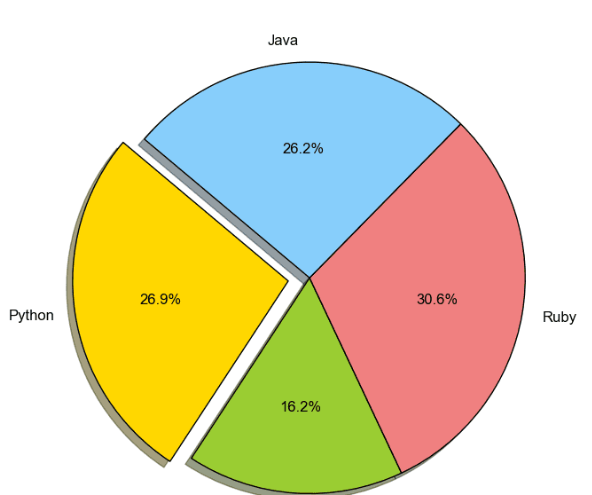
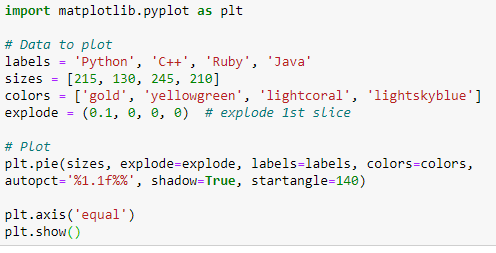
A scatter plot is a type of plot that shows the data as a collection of points. The position of a point depends on its two-dimensional value, where each value is a position on either the horizontal or vertical dimension



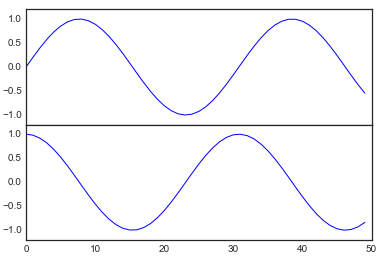
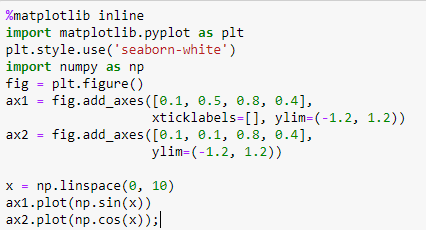
With the use of stack plots, The idea of stack plots is to show "parts to the whole" over time. A stack plot is basically like a pie-chart, only over time. All I did here was plot some empty lines, giving them the same colors, and the correct labels in accordance with our stack plot. I gave them a linewidth of 5, to make the lines a bit thicker in the legend. Now, we can easily see how we're spending our days!



Pie object is a circular statistical chart, which is divided into sectors to illustrate numerical proportion. Data visualized by the sectors of the pie is set in values. The sector labels are set in labels. The sector colors are set in marker.colors.



Sometimes it is helpful to compare different views of data side by side. To this end, Matplotlib has the concept of subplots: groups of smaller axes that can exist together within a single figure. These subplots might be insets, grids of plots, or other more complicated layouts.



**Summary/Conclusion**

Data visualization like the use of charts, line charts help to visualize data in a compact and precise format which makes it easy to rapidly scan information in order to understand trends. is important to design the right visualizations for your data to allow yourself and team members to interpret and make decisions based off of what they observe.

**References**

GargEntrepreneur, H. (2019, July 29). Excel Tutorial for Python and Pandas. Retrieved from <https://www.dataquest.io/blog/excel-and-pandas/>.

Matplotlib scatterplot. (2015, May 30). Retrieved from <https://pythonspot.com/matplotlib-scatterplot/>.

Matplotlib Tutorial: Python Plotting. (n.d.). Retrieved from <https://www.datacamp.com/community/tutorials/matplotlib-tutorial-python>.

PrykePython, B., FinTech, & Machina Capital. (2019, September 11). Jupyter Notebook for Beginners Tutorial. Retrieved from <https://www.dataquest.io/blog/jupyter-notebook-tutorial/>.