Final project proposal

1. **Individuals in the group.**

In this project, I plan to finish individually.

1. **Details about the problem**
2. **What the problem is.**

The task is to conduct explorative data analysis on the price of avocados. The problem is to figure out the potential patterns in avocados, like seasonal patterns or cyclical patterns. Additionally, related factors were also considered or evaluated to suggest the influence on the avocados’ price.

1. **Why the problem is interesting – refer to the literature.**

It is well-known that Millennials in America really love avocados, which already becomes a fashion among teenagers. It’s also well known that Millennials live with their parents, which means they do not have much freedom to buy as many avocados as possible. Clearly, they aren’t buying home because firstly the avocado was not cheap, secondly, they are supervised by their parents. So, maybe there’s hope for American Millennials to find a city with cheap avocados, they could achieve the “Avocado Freedom” to some extent and start a life there.

In addition, as an applicant in American master’s degree and a fan of avocado. I would take this factor into consideration. Therefore, it is of great importance to analysis the price trend and possibly make some prediction about the future avocados.

1. **Relevant work – refer to the literature.**
2. Previous works on the price of avocados prediction from 2015-2018

<https://www.kaggle.com/janiobachmann/price-of-avocados-pattern-recognition-analysis>

1. Prophet R package (API)

<https://facebook.github.io/prophet/docs/quick_start.html>

1. **Information about the data source:**

**A. What data you plan to use.**

- date: The date of the observation  
- average price: The average price of a single  
- total volume: Total number of avocados sold  
- year: The year  
- type: conventional or organic  
- geography: The city or region of the observation

**B. Where you plan to get it from.**

The datasets represent weekly retail scan data for National retail volume (units) and price. The table contains the data from 4 January 2015 up to 17 May 2020. Retail scan data comes directly from retailers’ cash registers based on actual retail sales of Hass avocados. The Average Price (of avocados) in the table reflects a per unit (per avocado) cost, even when multiple units (avocados) are sold in bags. The Product Lookup codes (PLU’s) in the table are only for Hass avocados. Other varieties of avocados (e.g., greenskins) are not included in this table. I downloaded the dataset from the Kaggle. <https://www.kaggle.com/timmate/avocado-prices-2020>

1. **Proposed methodology (including subtasks, methods used in the analysis, tools, processes etc.).**
2. Compares two types of avocados: conventional and organic by considering the average price and time series.
3. Find the potential relationship or patterns between average price and volumes per year
4. Analyzing the seasonal patterns, finding out the peak and drop of avocado price in each month, which is essential to make prediction of the price in time series
5. Find the place with cheapest avocados
6. Use the prophet R package to forecast the future price.
7. **Final evaluation methods and criteria.**
8. we could make a comparison between the previous dataset prediction and the actual price
9. I need to make a summary list to conclude the research findings and make potential prediction.
10. Use machine learning models to make the prediction and evaluated with ROC or PR curve (optional)

**6. Potential limitations and challenges.**

1. Use the Facebook API prophet r-package may be not helpful for understanding the background mechanisms.
2. I could not filter the cities and grouped into 4 clusters, like west, middle, east because I am not American, many cities were not really familiar to me. It is better to divide the region based on some broader scale.