Hello everyone. Thank you for the time in listening my presentation. Today, my topic is about the analysis of avocados in the United States, I would conduct both explorative and predictive data analysis from 2015 to 2020. Specifically, the whole work was finished by R.

This is the overview of my whole presentation. It composed of three parts: data preprocessing, explorative data analysis and predictive data analysis. I would mainly focus on the EDA part, including the comparison between two types avocados, the association analysis of price and sold volumes, seasonal patterns analysis of avocado, the geographical analysis to return the cheapest region in the united states. Meanwhile, the future price trend would be predicted by prophet r package to conduct the future analysis.

Now, let’s move to the first part, data preprocessing part. I returned the head of dataset to have an overview of inside components. In this case, I would use six features in the dataset, average price, total volume, year, type and geography. Furthermore, it is significant to explore the number of missing values, the NAs. From this figure, we could see that the dataset I used do not contain any missing value. So, I could move on to the explorative data analysis, which is the most important part of my research.

To figure out the difference of the avocado types, I firstly used the density plot to compare the average price distribution. As can be seen from the plot, the green stands for the conventional while orange stands for the organic. The overall shape of conventional type is right skewed shape and the organic type is relative normally distributed. In other words, the conventional avocado tends to center around the relative low price. Contrarily, the organic type distributed more evenly, the center could be around 1.5$. Meanwhile, for better illustration of the price trend in these two types, I plotted the avocado price in last five years. It is clear that the organic type always has a higher price compared to the conventional and the price for both types seem to follow some seasonal patterns, which required me to explore it further in the following part. Based on previous research, I made a table to summarize the results. As can be seen from the table, conventional avocado could account for approximately 97% in the markets and the average price would be 28% lower than organic type.

It is a common view that the price and volume sold obeys some correlation. We would like to see whether the avocado follows this relationship. Apart from that, the previously mentioned seasonal patterns are also my focus. Hence, I visualized the average price and volume of both organic and conventional avocado in each month. Figure A, B stands for the average price in each month. Blue dashed line stands for the minimum value while the red dashed value stands for the max value. The most expensive conventional avocados in one month could 1.8$ and the cheapest month of conventional avocados can be 0.82 $. Similarly, the most expensive organic avocados in one month could 2.1, the cheapest month of conventional avocados can be 1.21. Figure C, D stands for the volume sold in each month, the red line stands for the trend. From the figure we could see that the love for conventional avocados of American people is consistent, which could be reflected by the continuously grown sold volume. Regarding with the organic avocados, in 2019-2020, possibly owing to the decline of economic situation and the covid 19 pandemic after 2020, the trend total volume sold started to decrease. From the bar chart, it is clear that the avocado follows some seasonal patterns, a common feature shared by the fruit. (这里需要前后页ppt翻)

Let’s move to the next part, the analysis in each season. I compared the average price, volume in each season in last five years. Figure A, B stands for the average price of avocados for either conventional or organic and Figure C, D stands for the volume sold for either conventional or organic, the unit is million (m). Since conventional avocados account for almost 97% of the markets, I would mainly focus on the conventional type. I used the red grid to indicate the maximum volume sold season in each year. From the figure C we could see that, except in 2017, people mostly buy avocados in spring and summer. Additionally, I used the blue grid to indicate the season with lowest avocado price in last five years, either winter or spring could be the lowest season. We could find that winter and spring are connected, same with spring and summer. The possible reason for this situation could be the month difference just between these two seasons. As a result, I would analyze both price and volume sold in each month of the last five years later. Furthermore, the results follow the common sense, volume sold increase as the price decrease and of course it would have some delay reflected in the volume sold.

Here is the line chart of price and volume in each year. The x axis indicates the month of the year. Again, I would focus on conventional type, which is representative. We could see that the cheapest price consistently come in February, while the highest price months are relatively arbitrary. Regarding with the avocado sold, generally, there are two peaks in both February and May and the lowest month is not restricted. Moreover, I combined the past five year’s data into one line chart for more detailed representation, elucidating the previous arbitrariness. We could see that generally, the peak for average price could come in September and lowest come in February. For avocado volumes sold, both February and May are two peaks of the whole year while people in November show minimum willingness in buying avocados. Based on the research, the avocados tend to ripe in August and September, plus the time in transportation and packaging, it is unavoidable that it could have about one month late reflected in the price. Overall, the price trend and avocado volumes trend are negative correlated, which obeys the common law. I made a summary of the previous figures; you could stop if you would like to see.

Next part would be the geographical analysis in US. I returned six regions with lowest price for two types avocados. Again, I would focus on the conventional type. The lowest average price could be 0.78 $, almost a half of the organic types. I marked the location of these region on Google maps, we could see that these regions are approximately in same latitude and near Mexico, which is the origin of avocados. Therefore, I could hypothesize that these regions are the center of avocado plantings and therefore such low price in avocados.

Next part would be predictive data analysis. In this case, I utilized the Facebook API r package to predict the future price trend. From the plot, we could see that the price would continue to follow the same patterns of previous years. And this plot was a more detailed price trend yearly, monthly, weekly and daily. If the value in y axis is positive, then it suggests the price is increasing. Figure A is the yearly analysis. Since 2015, the average price of avocados continues to grow and 2017 was a crazy year, they grow in a speed higher than 1.3; After 2018, the average price tends to be stable, but still has a rate about 1.28; Regarding to the future prediction, the range is about [1.20-1.34]. But considering the existing pandemic in North America, the possible rate could be 1.20. These slides are the summary of previous four figures. Stop if you wish to see the points.

That’s all for my presentation. Thank you for listening.