Pork Price Analysis Report

1. Problem Statement & Data Description

Pork is the mostly consumed source of meat in China, which is about three times as much as the runner up poultry. The price of pork is an important index of pig-raising industry and it's also highly correlated with wellbeing of people's lives. For this particular problem, we are interested in the price variation trending of pork during a six-week time period from September 21st to November 2nd.

We make a price recording every two weeks, and for each recording, it consists of the **time**, **city**, **the name of pig's part** (could be loin, ham, ribs…), **market type** (wet market or supermarket) and of course the detailed **price** of pork. So the dataset has 5 dimensions in total, and we are trying to figure out the overall price trending along with the details and distinctions between different cities, pig parts and market types.

2. Data Wrangling and Cleansing

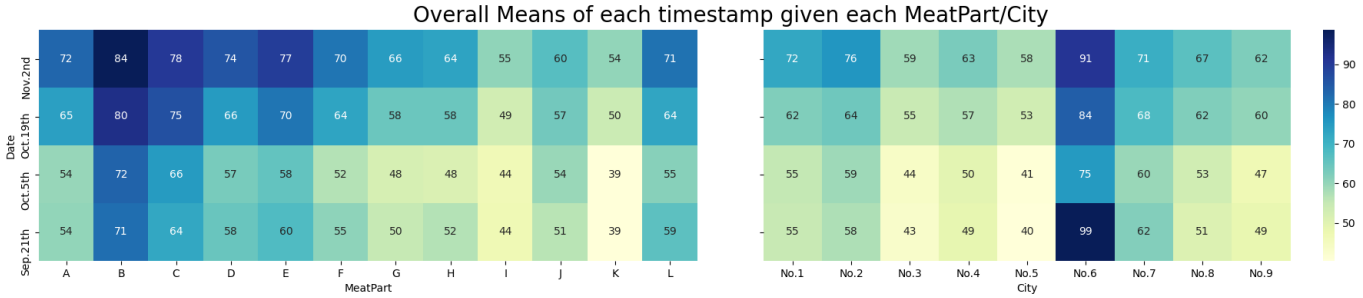
The original data are recorded as excel sheets, and I prefer to use python pandas to wrangle them so that in later analysis I can easily group them. To be more specific, I create a data frame with five columns: 'Date', 'City', 'MeatPart', 'Market' and 'Price', each of them represents a dimension. Thus, a row of this data frame would contain all the information for a specific record. The first four variables are categorical and the Price is numeric.

Then we want to do some data cleansing by checking typos and dealing with blank. After checking the unique values for each categorical variable and the range of Price, I think there is no typo/error in this dataset. The blank recordings only exist in Price attribute, and among all the 984 Price recordings, 48 of them are missing. So to handle such situation, I build a Random Forrest regression model based on the non-blank data which takes four categorical variables as inputs and Price as output. When applied to test dataset, The RMSE (root-mean-square error) is about 6.6, which is acceptable to me. Thus I use this model to predict the missing Price values given their Date, City, MeatPart and Market type. This dataset will be used for later analysis and is store as excel file 'Wrangled_data_pork_rice.xlsx'.

3. Data Visualization

To get a better intuitive understanding of data, I also did some plotting before heading into statistical analysis.

(1) First, I'd like to compare the prices among different values of dimension 'MeatPart' and 'City', along with the time stamp. Thus under such demand, I use two heatmaps as Figure_one with each cell represent the mean of all the price recordings that satisfy its coordinate.



Figure_one

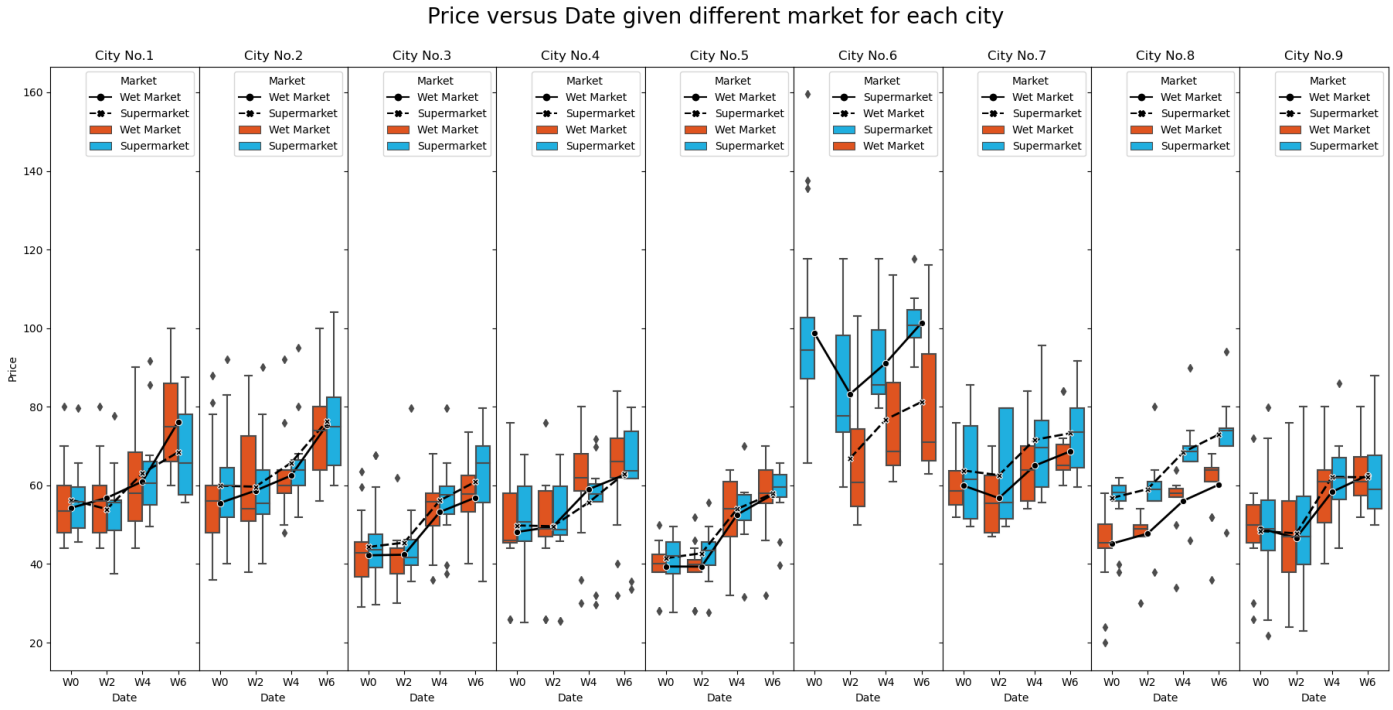
From the above two heatmaps, we can make following inferences:

① If we focus on the vertical axis, then it's obvious that the prices of all categories are rising from September 21st to November 2nd, especially after October 5th. Besides, city No.6 seems special compare with other cities. Instead of a steady rising-up pattern, its pattern is more like a roller coaster, and the prices are significantly higher than other cities.

② If we look it horizontally, then we can see that the relative order of prices among there categories remains stable. For example, part B is always the most expansive part and part K is always the cheapest. City No.6

always has the highest price while city No.5 lowest. Such phenomenon means that the rising pattern may not be regional, it could be a national, wide ranged problem.

(2) Then I want to add the Market variable by making two boxplots Figure_two and Figure_three, one for different cities and the other for different parts of pork.



Figure_two

We can see that the prices from wet market and supermarket are highly correlated, which is reasonable. It seems that prices of pork sold in supermarket is a bit more expensive than those sold in wet market, and there is no obvious pattern about the variance of these prices.

Figure_three

From Figure_three we are more confirmed about the higher prices of supermarket pork. However, there seems to be also a subtler difference between wet market and supermarket: the prices of wet market rise more in first two weeks than those of supermarket. I would like to verify this later.

4. Detailed Analysis and Results

To get detailed analysis statistics on the price variation trending both qualitatively and quantitatively, I generate two sheets to record these statistics, one for dimension City and another for dimension Part of pork. For each table, it consists of twelve columns representing as follows:
(1)The name of Part/City index | (2)The minimum of mean of pork group by Date | (3)The maximum of mean of pork group by Date | (4)The Date that reaches minimum mean | (5)The Date that reaches maximum mean | (6)The overall rise | (7)The Date that reaches minimum standard deviation | (8)The date that reaches maximum sd | (9)The overall rise for wet market| (10)The overall rise for supermarket | (11) The rise between W0 and W2[1] for wet market | (12) The rise between W0 and W2 for supermarket

All the statistics calculated are saved in file 'analysis_statistics.xlsx'.

5. Conclusions and Discussions

(1) Conclusions from statistics:
① all the prices of different parts of pork and different cities (except city No.6) reach their peaks on the last time stamp November 2nd, resulting the overall price rise for these categories ranging from 16% to 42%.
② give the data we have, there is no obvious pattern of the standard deviation of prices.
③ the supermarket has a lower price rise both in overall six weeks periods and in first two weeks, especially for part A, B, C, D, J, K, L and city No.1 and No.2. This is probably because compare with wet markets, supermarkets usually have more stable supply of product and also more abundant stock.
④ As mentioned before, city No.6 has different price trending compared with other cities, which requires further investigation.
(2) Discussions and future work
① For this analysis, I use regression model to fill in blank data, due to limited time, I don't use cross validation to do more inspections of my model, and more models can be tested. We can also try to delete these blank data, it may result in different conclusions.
② Combined with Figure_one, the middle of October could be an important time stamp because prices start to rise significantly, and we could dig out more reasons for that by searching for more information. For example, if this data is from 2019, then the outbreak of African Swine Fever should be considered.
③ One of the major problems of this dataset is its short time period. It's not convenient to do time series analysis given only four timestamps thus we need to trace longer time period, and collect data from more markets for each city so that we can dig deeper into part of meat/city dimensions.

[1] W0 for Week 0, September 21st, W2 for Week2, October 5th