FIT 1043 Assignment 2
27716503_Lee Yee Voon

Part A

Task 1:

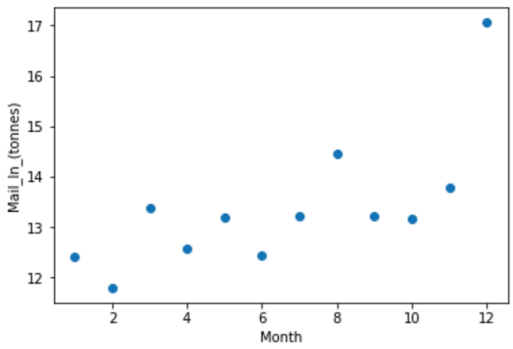


Figure 1

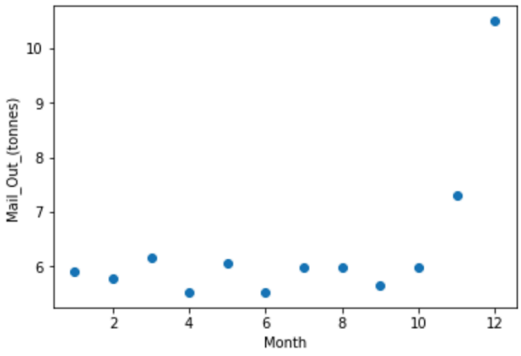


Figure 2

1. **Which month of the year are the most mail packages sent both to and from Australia? Why do you think that is?**

   December has the most mail packages sent both to and from Australia.

2. **Overall, do we see more mail volume entering Australia from the US or leaving Australia for the US?**

   As shown in Figure1 and Figure 2, there are more mails entering Australia instead of leaving Australia. Hence, there is more volume entering Australia from the US.

Task 2:

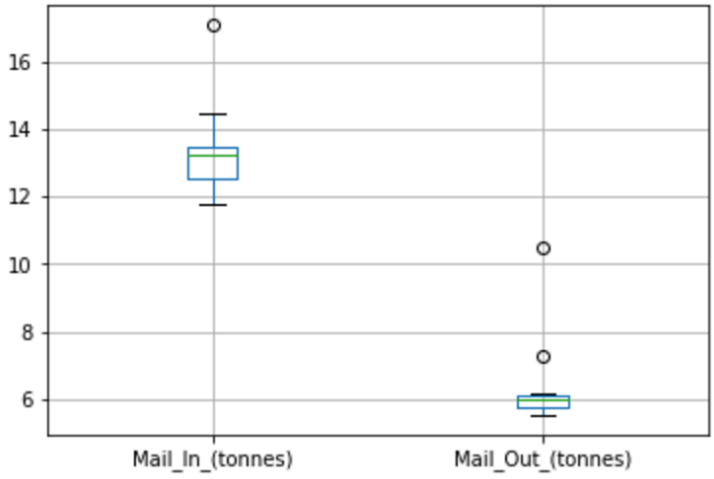


Figure 3

The boxplot for Mail_In_(tonnes) and Mail_Out_(tonnes) have been plotted together.

1. **How many outlier values can you see in each boxplot?**

   There is 1 outlier value in the boxplot with “Mail_In_(tonnes)”. There are 2 outlier values in the boxplot with “Mail_Out_(tonnes)”.

2. **By looking into each boxplot and its relevant scatterplot, which month of the year do the outliers in each boxplot correspond to?**

   For "Mail_In_(tonnes)" boxplot, it is the month December. For "Mail_Out_(tonnes)" boxplot, it is month November and December.

3. **Delete outliers and generate the boxplots again.**

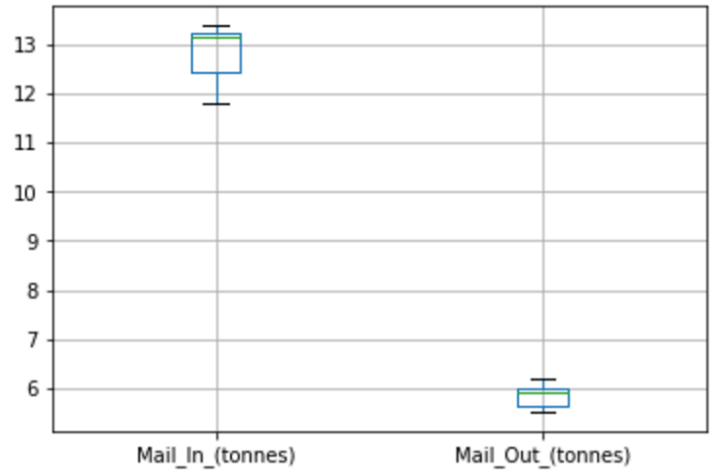


Figure 4

Part B

1. **Does the data show a clear trend? If you don't see one, try plotting the data after aggregating at the year level.**

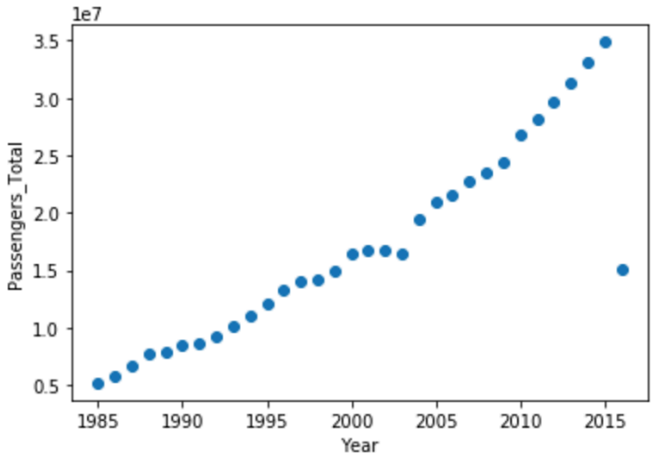


Figure 5

   Yes, the data shows a clear trend of increasing in total number of passengers as the years goes by.

2. **Is there a problem with one of the data points? If so, why is that?**

   Yes, as shown in Figure 5 there is one outlier in year 2016. This could be due to incomplete data sets, because there are only 5 months' worth of data for 2016.

3. **Remove the problematic data point and run a simple linear regression in Python (or R) by modifying the code from the tutorial. Does the linear fit look to be a good fit to you?**

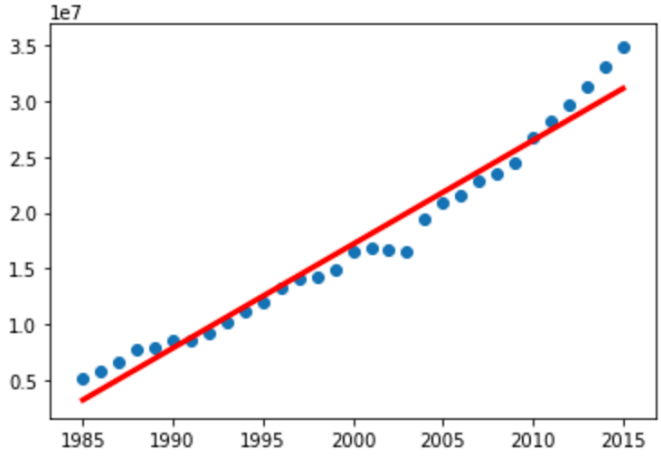


Figure 6

Yes, it intercepts more data points.

4. **How fast are international passenger numbers increasing each year? [Hint: What is the slope of the linear fit above?]**

The linear regression line has a slope of 928503.0, this shows a massive increase in rate of change

5. **What does the linear model predict for passenger volume in 2020? [HINT: Get the slope (m) and intercept term (c) for the linear fit above and use the function Y=m*X + c, to calculate the prediction for X=2020.]**

Y = m*X + c
Y = 928503.0 * 2020 + (-1839827126.967742)
= 35748933.032

Hence, the total number of passenger volume predicted in 2020 is 35,748,933.

6. **Try fitting the linear model only to the data from the year 2004 onwards. What happens to the prediction for 2020? Which prediction do you trust more? Why?**

Y = m*X + c
Y = 1399806.751748 * 2020 + (-2786554191.7214451)
= 41055446.809515

Hence, the total number of passenger volume predicted in 2020 is 41,055,447.
This prediction should be more accurate, due to the timeliness of data collected. The trend may differ for different years and it is more accurate to predict based on more up-to-date data, which in this case is the data from the year 2004 onwards instead of from the year 1985.

Part C

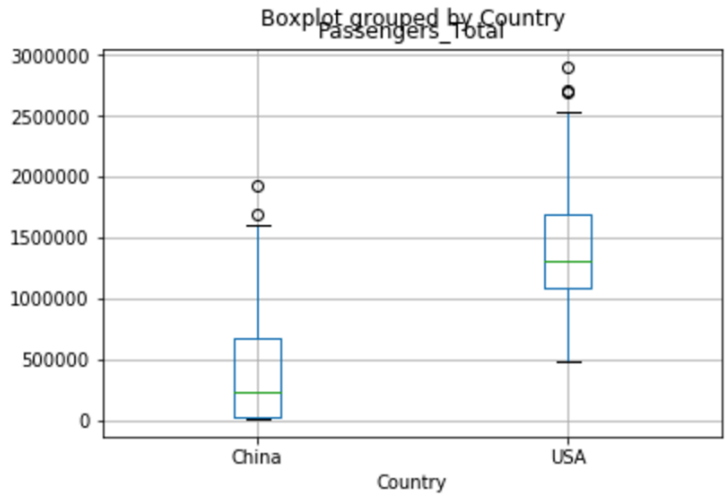


Figure 7

1. **By looking at the median values in boxplots, which country has a higher median total number of passengers?**

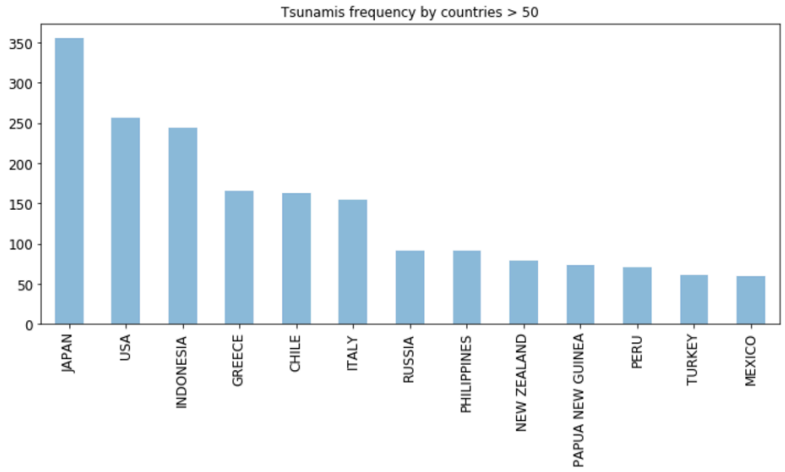
   USA has a higher median total number of passengers.

2. **What is the difference between the minimum values in USA and China boxplots?**

   Min. of China: 0; Min. of USA: 500,000
   Difference = 500,000-0 = 500,000

Part D

1. Data was aggregated based on the frequency of tsunami incidents for different countries and for year 2000 - 2017. It has shown that the highest frequency of tsunami occurrence is Japan, and lowest frequency of tsunami occurrence is Mexico.



2. A boxplot is plotted to compare the magnitude of both of the countries.

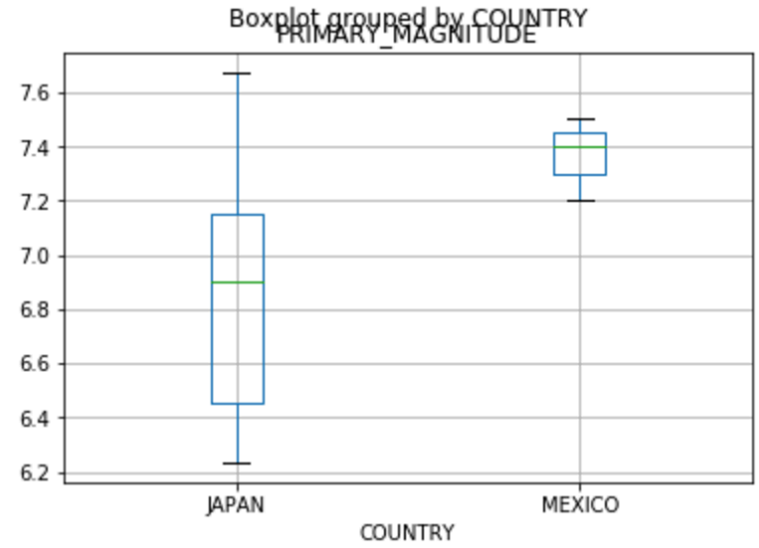


Figure 8

According to figure 8, the difference between the minimum value for Japan and Mexico 1.0. Due to the data from recent years, there were no outlier for both countries. However, if data from 1900 to 2017 was used, there would be an outlier for Japan. As shown in Figure 9 below.