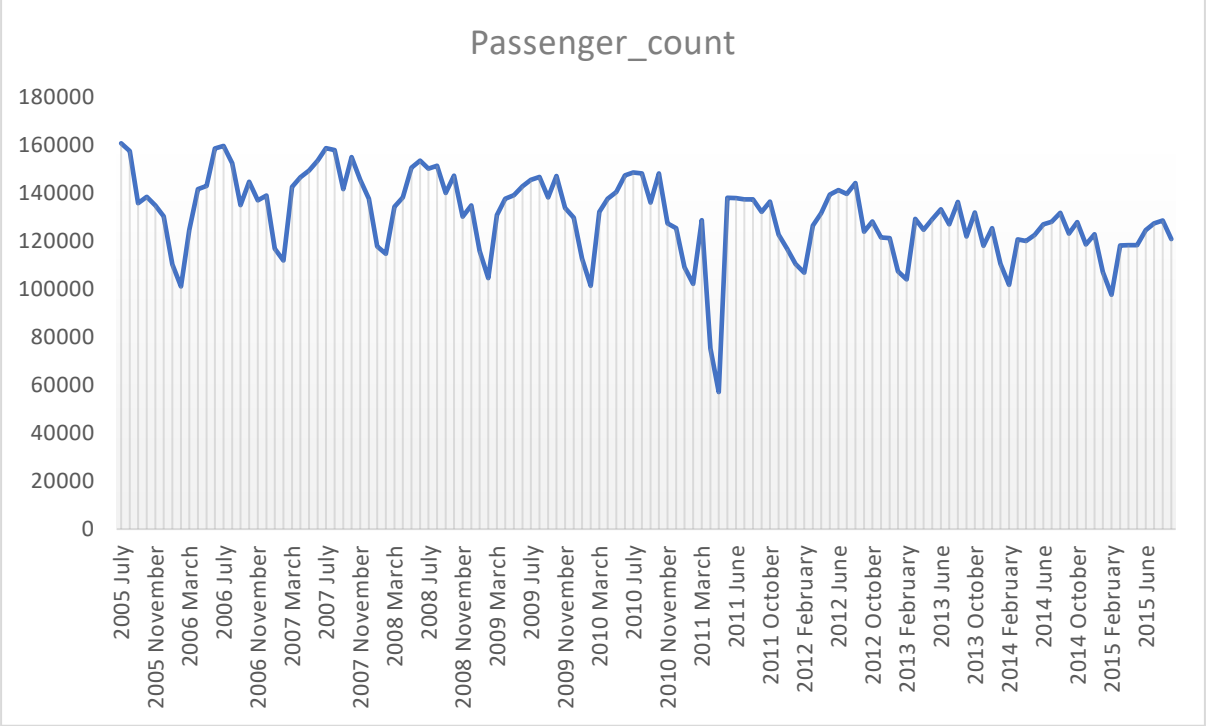
DS250 Exam 3

1. [30 marks] Below are the monthly counts of passengers travelling by American Airlines on domestic flights in America from 2005 July to 2015 September. You are asked to analyze the data and answer a few questions.



Month	Passenger Count	Trend	Seasonality	Noise
2005 July	160890			
2005 Nov	135008			
2006 Mar	124880			
2006 July	159845			
2006 Nov	137123			
2007 Mar	142766			
2007 July	158846			
2009 July	145690			
2011 July	137558			
2013 July	127146			
2015 Oct				
2015 Nov				
2015 Dec				
2016 Jan				
2016 Feb				
2016 Mar				

[5] Which month(s) in the above data seem to be Anomaly? Please remove them in your analysis.

[25] Decompose the number of passengers into Trend + Seasonality + Noise terms. Write equations for each of the terms and complete the table above.

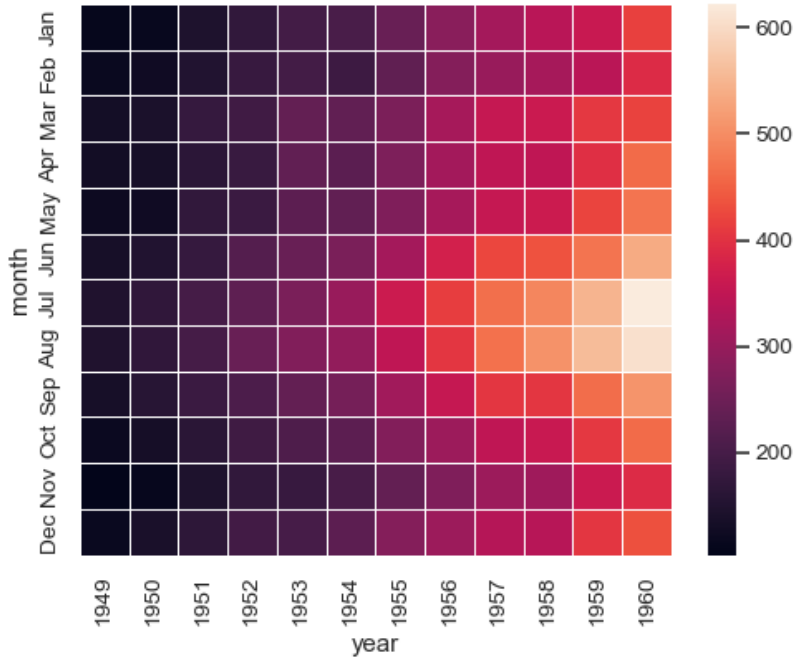
2. [30 marks] Ajay was given a huge dataset of hundreds of millions of images by his manager and asked to organize them in groups. As a first step, he used Resnet-50, a popular ML model to extract features from the images. His feature vectors are of dimension 150.
   [5 marks] What are some difficulties he may face in clustering the data? Which steps shall he take now?

   [10 marks] Assume, the data has been converted to 2 dimensions and the following points are obtained. Perform a clustering of the following data using K-means++ with K=3. Justify your steps.
   Data Points: (2,3); (3,2); (4,1); (4,4); (5,1); (5,5); (5,6); (6,6); (1,5); (1,4); (5,2); (1,7)

   [10 marks] Compute the stats required by the BFR to summarize each of the clusters. Now add the points (2,2); (4,3) and (3,4) to the clusters using BFR algorithms. Which cluster does each point go to?

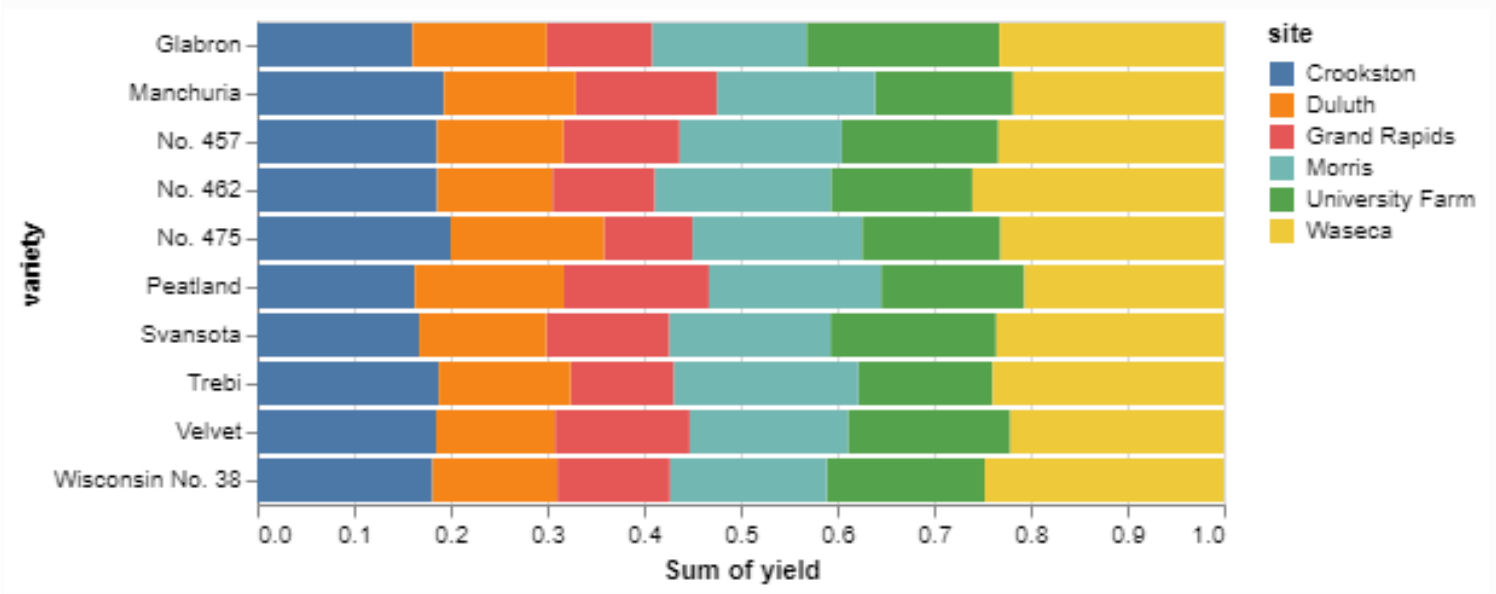
   [5 marks] If we use DB-SCAN to cluster these points (including the newly added points), how many clusters will we get? Justify.

Q3a. [5 marks] This is a heatmap plot of passengers who took flights during 1949 to 1960.

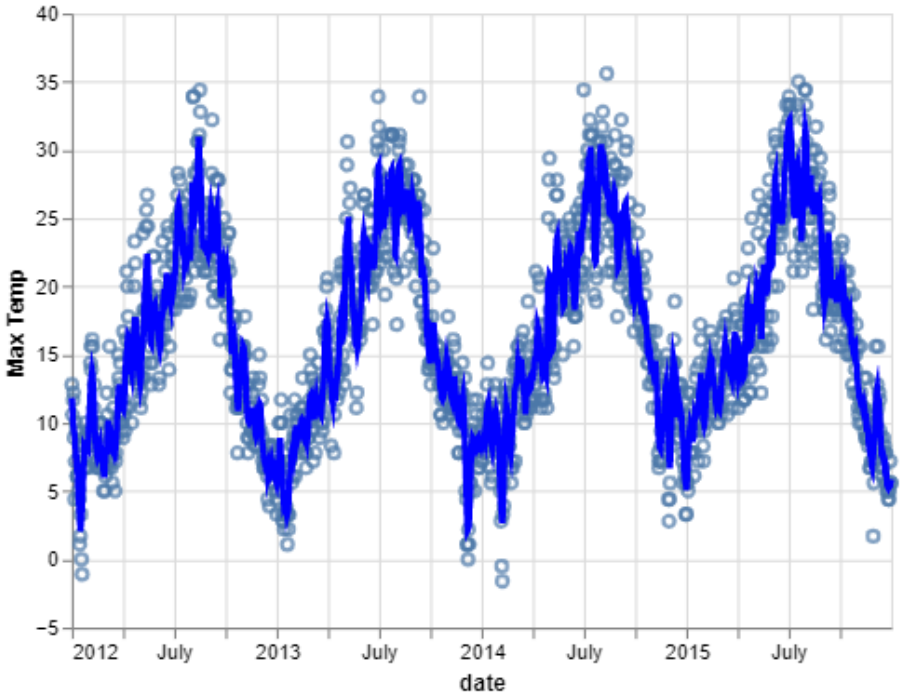


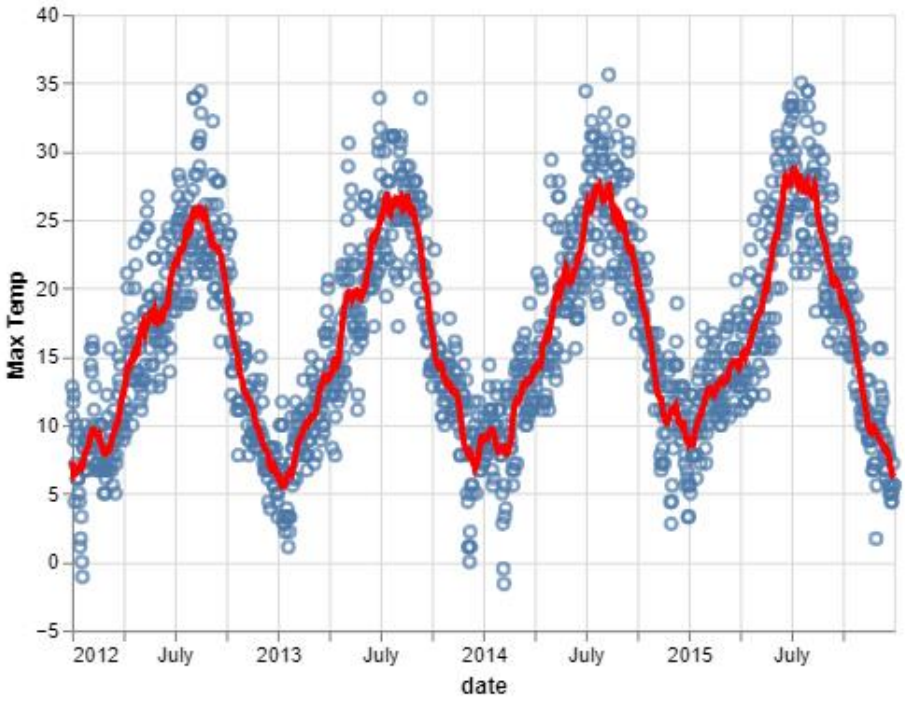
Identify the busiest month in this whole period.

Q3b. [5 marks] This is a normalized stacked bar chart using data which contains crop yields over different regions and different years in the 1930s. Which site has the least production of No. 475 variety?

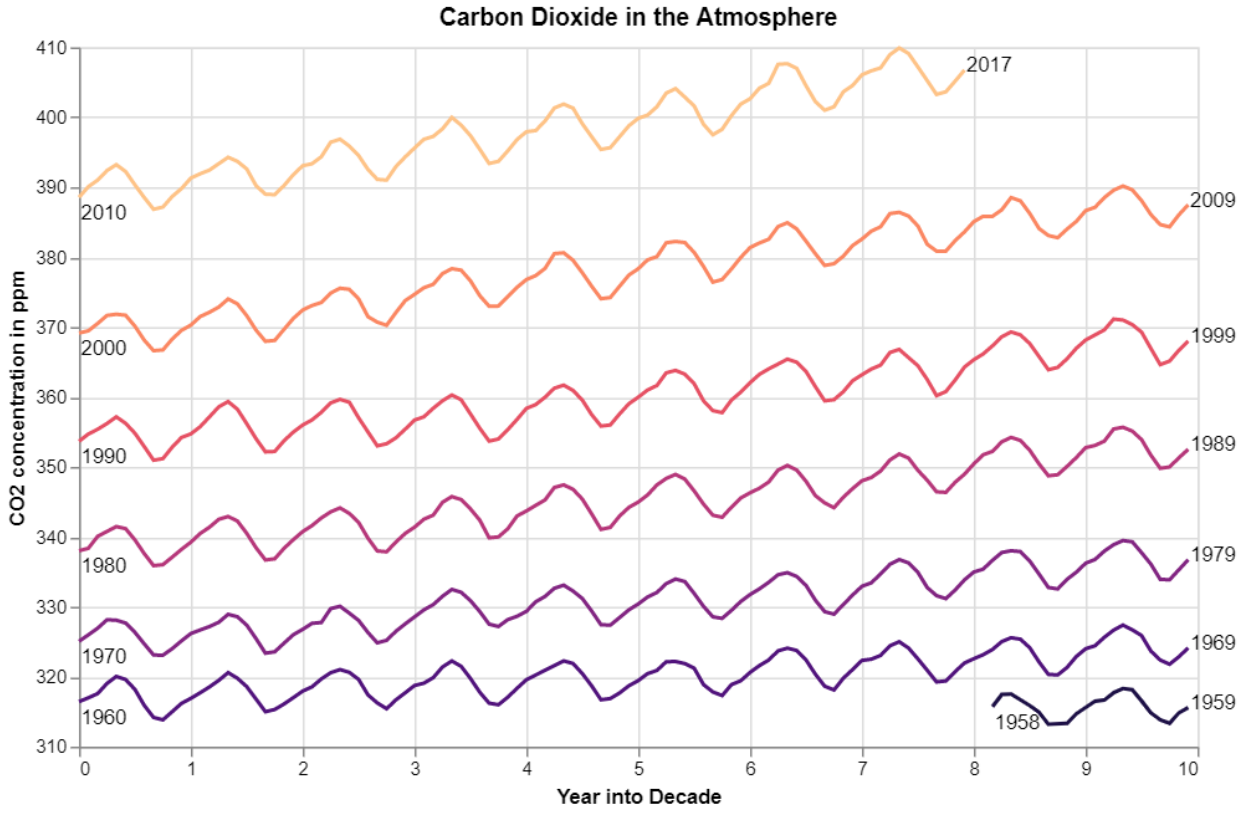


Q 3c [5 marks] The max temperature in the city of Seattle is plotted here as a scatter-plot with Rolling mean. There are two plots here, one with rolling mean window of 30 days and another with window of 5 days. Identify them and write what they each represent?





Q3d. [5 marks] This example is a fully developed line chart that uses a window transformation. Write a brief interpretation of this chart (no more than 50 words).



Q3e: [10 marks] Can you identify the anomaly in this Trellis plot regarding the yield of barley in years 1931 and 1932 in different regions/sites?