1. [15 points] Assume that you are working for a credit card company. Recently something surprising has come across to your attention.
   - Like most of the credit card companies or banks, your company has been trying to encourage customers to sign up for e-statements. Once customers agree to e-statement, your company can save the cost of mailing paper statement to customers. In the past, this has been viewed as a practice without introducing any credit risk. In fact, some people tend to believe that credit-savvy customers are more likely to sign up for e-statements.
   - Recently, however, someone did some analysis that led to surprising result. Based on this analysis, the e-statement customers have higher likelihood to charge off (default) on the credit card. As shown in the chart below, by examining the number of charge offs in 2015 over the number of open accounts at the beginning of 2015, those using e-statements have 33% of default rate versus 19% among those who are not on e-statement.
   - Once you can download the Credit Performance Data file from Canvas, you will have both the data and the analysis performed. Just a little more on the data:
     - This data captures the account status by month since the account opening
     - Most of the accounts would start as paper statement and then some move to e-statements, where there would be "Yes" in the e-statement column
     - Customer account would be open, closed, or charge-off in most of the cases. When customers start missing payment, there could be 30, 60, or 90-day delinquency status on the account.
     - The "credit limit" column shows the limit on the credit card
     - The "balance" column indicates the amount of balance on the card. When people close an account, they should have zero balance. When people default on their cards, most of them have balance to their limits.

   Please leverage this data and descriptive analytics to address the following questions
   a) Do you have any concerns on the suggested conclusion based on this analysis?

   Following are the concerns related to the given analysis:
   - Probability of defaulters is calculated from the open cases for the month of Jan. It seems incorrect to compare charge off cases for the whole year against the open cases for the month of Jan. It could very well be that the cases were open from long before. Perhaps more data is needed to draw meaningful conclusions.
   - There are cases where an account is closed but the balance is present. Ideally, close account should not have balance. These accounts have not been included in the analysis.
   - Different years presented us with different analysis for e-statement users. Prior to 2015, user accounts that did not use e-statement had more probability for defaulting. 2015 shows a contradictory story.

- Also, other factors could contribute to the credit risk analysis like Balance and credit limit which have not been fully explored.

b) List other findings that will either support or contradict the above analysis
- If same analysis is produced for the Jan 2014 open cases against 2014 charge offs, the probability contrasts with the 2015 analysis. Probability of defaulters is higher for those accounts which haven't switched to e-statement.

  This weakens the given analysis.
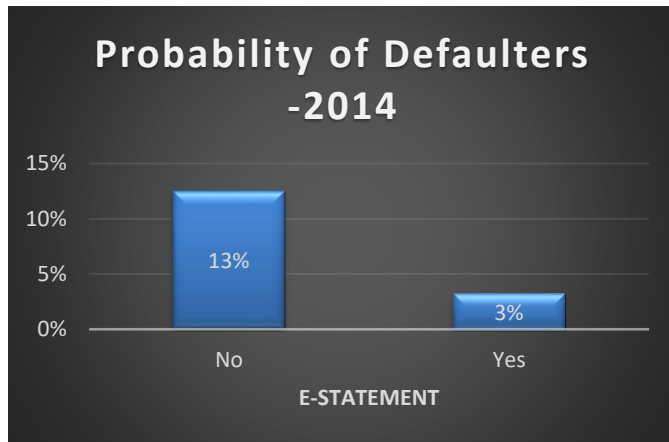
- Three closed accounts 10000120, 10000220 and 10001006 where balance is greater than $0 have not been accounted for in the analysis.
- Out of the accounts that charged off, the users that used e-statement had a greater balance associated. Implies e-statement users tend to charge off more money.

c) What's your assessment on the relationship between e-statement and credit risk and provide a single chart to support your claim.
There is no strong relationship between e-statement and credit risk. In total there are 70 accounts provided to us. 24 out of 70 accounts have been identified as charged off. There is an equal distribution of 12 each for users which use e statement, and which don't. The graph below depicts there is no significant trend available between e-statement users and their charge off distributions to draw conclusions from.

Charge off distribution
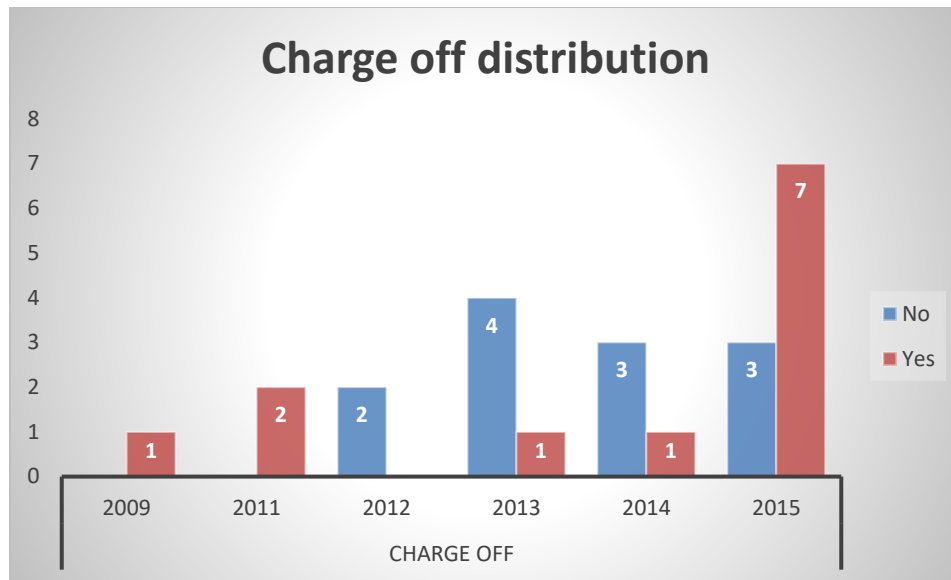
2. [10 points] Download the "City of Chicago – Homicide" dataset from Canvas. Use this dataset to answer the following questions:

a) Suppose that the Chicago Police Department is concerned about lack of resources in some districts, particularly among those seeing sharp increase of homicides and significant decline of homicide arrest rates (defined by the number of arrests over the number of homicides) between 2015 and 2016. Which three districts are the most concerning districts?
To find the most concerning districts, we are looking to maximize the number of homicides and minimize the homicide arrest rates.

District	Increase in Homicide (between 2015 and 2016)	% Incr in Homicide ((between 2015 and 2016)	Decline in Arrest
1	4	80%	29%
2	7	30%	17%
3	14	56%	-13%
4	11	38%	7%
5	14	47%	-14%
6	17	41%	-9%
7	46	118%	-8%
8	28	93%	-14%
9	34	106%	-4%
10	19	63%	-22%
11	44	90%	-14%
12	7	37%	-20%
14	-2	-17%	-48%
15	28	88%	-13%

16	-1	-20%	-30%
17	-3	-23%	4%
18	4	400%	-80%
19	-1	-11%	-19%
20	-4	-57%	-57%
22	3	14%	-33%
24	2	20%	-17%
25	5	15%	-7%

% Incr in Homicide is calculated as increase in homicide from 2015 to 2016.

Decline in Arrest is defined by the number of arrests over the number of homicides.

Clearly Districts **7,8** and **11** have experienced maximum increase in homicides at the same time the arrest rate has suffered a decline. Although between district 8 & 9 parameters draw quite close in but district 8 has a greater decline in arrest rates.
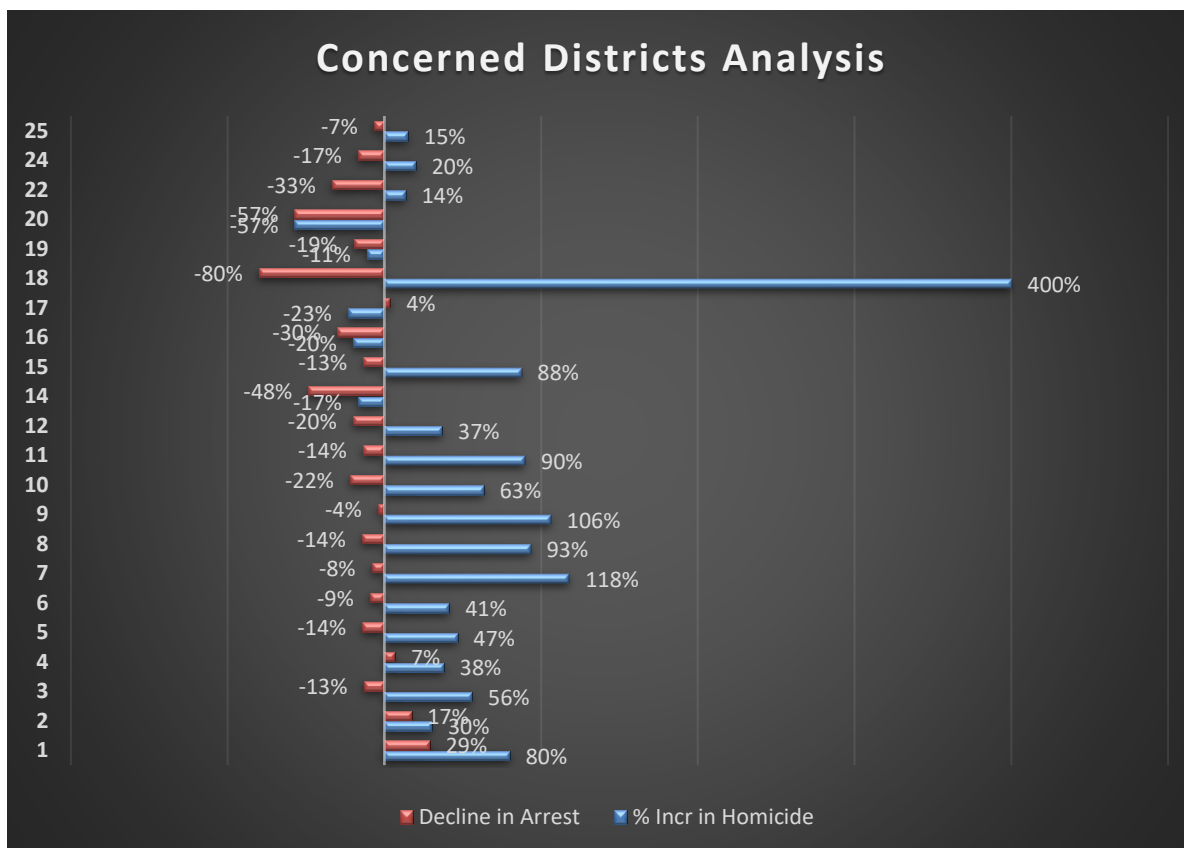
b) Assume that the Chicago Police Department also experiences resource constraints and it cannot add any more detectives. Instead, it plans to those districts with resource issues as identified in b) to share detectives with another adjacent district. This type of sharing should be 1:1 relationship. (For example, District 19 and District 20 are adjacent. If District 19 were to be identified in a) and District 20 was not,

then these two districts could share detectives. At the same time, District 20 could not share its resource with another district, and District 19 would not pull help from any other district.) How would you propose such a pairing solution to a)? Illustrate this through appropriate visualization.

Following image depicts the neighboring districts:
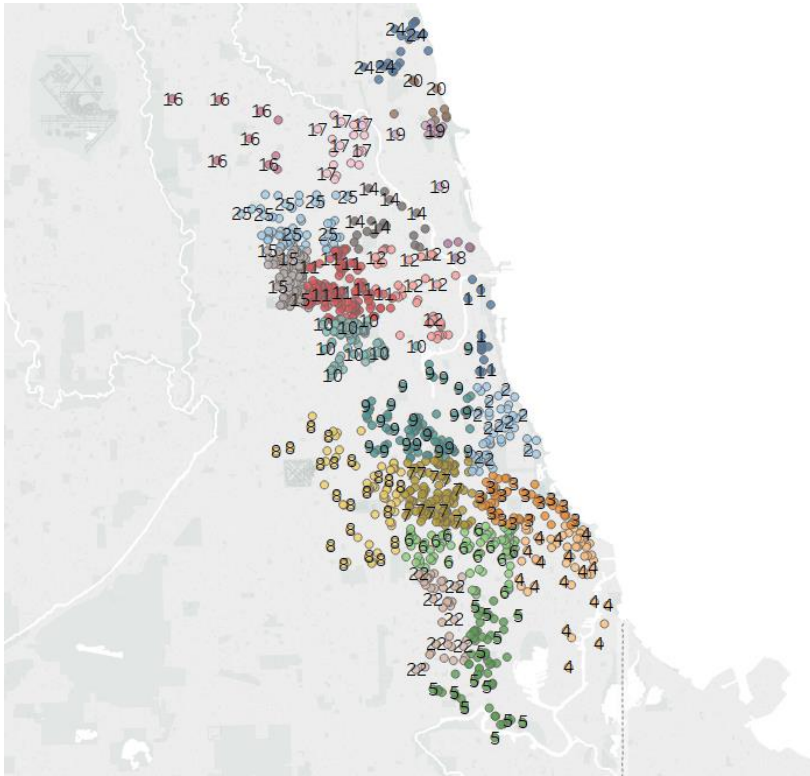
District 7 has districts 8, 9,2, 3 and 6 adjacent to it. To find an optimal pair, we would look for a district with relatively lower increase in homicide and higher arrest rate which would eventually provide desired support to district 7(in this case district 2).

After careful analysis of the adjacent districts homicide & arrest rates following are the proposed mappings:

High Alert districts identified			Proposed Pair		
District #	% Incr in Homicide	Decline in Arrest	District #	% Incr in Homicide	Decline in Arrest
7	118%	-8%	2	30%	17%
11	90%	-14%	25	15%	-7%
8	93%	-14%	9	106%	-4%