Unsupervised Learning (K-means)

Objective:

In this project, the main objective is to implement K_means clustering for the given data using two given strategies. They are,

- Strategy I: Picking the initial centroids randomly from the given samples.
- Strategy II: The first centroid is picked randomly and for the i-th center (i>1), the sample is chosen such that the average distance pf the chosen one to all the previous centroids is maximal.

Strategy I:

The goal of strategy-i is to randomly initiate the centroids from the given samples. To implement this library 'random' is imported. As the K-means clustering is to be done for k values of 2-11, a for loop is implemented for the overall code. From the obtained initial random centroids now, all the samples are clustered such that they are nearest to the particular centroid. Now the centroids and the corresponding clusters are obtained. The next step is to find the new centroids of the clusters by calculating the mean of all the samples in the cluster. Now, the new centroids are obtained, again the above process is repeated to obtain new clusters and centroids. This goes until the newly obtained centroids and the centroids in the previous iteration are the same(i.e the centroids don't change further). Now as we got the final centroids the objective function is calculated by the formula

$$Objective\,function = \sum_{i=1}^{k} \sum_{x \epsilon D_i} ||x - \mu_i||^2$$

The above steps are mentioned for all k values from 2-11

Results for Strategy I:

The given data is 2-d data with 300 samples, they are plotted as shown below.

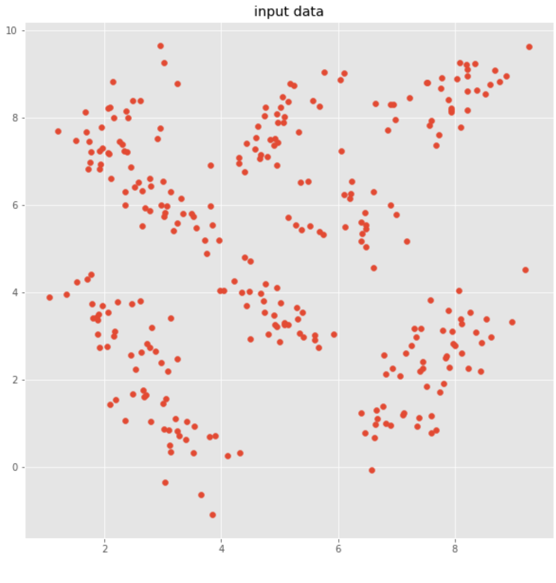
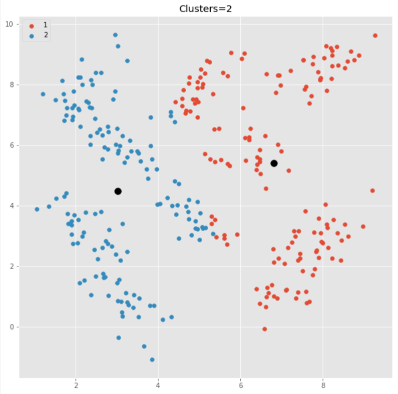
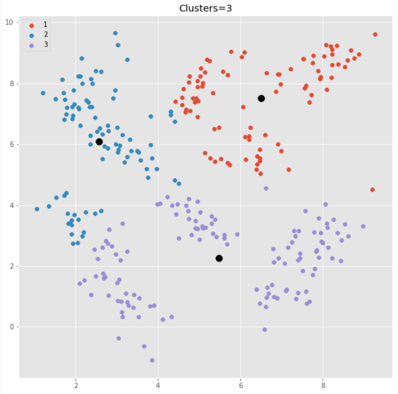
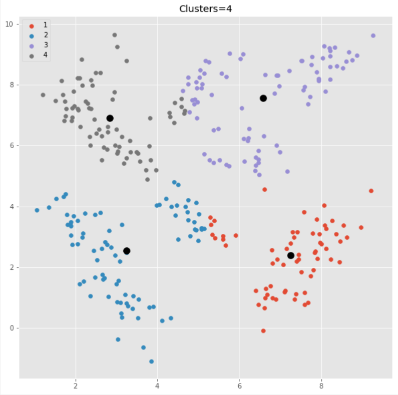
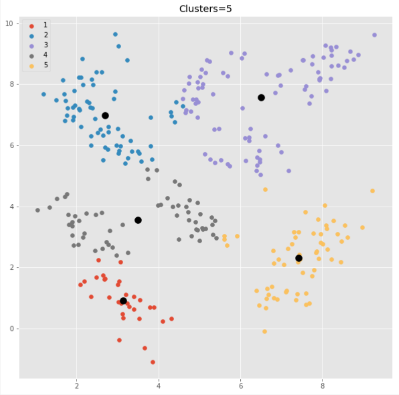
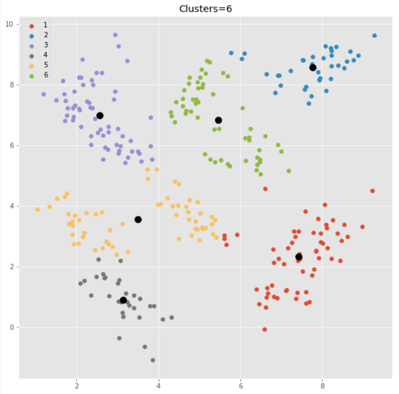
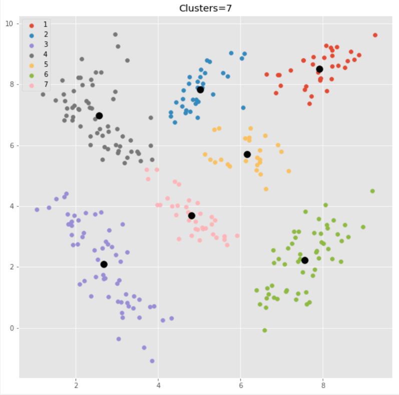
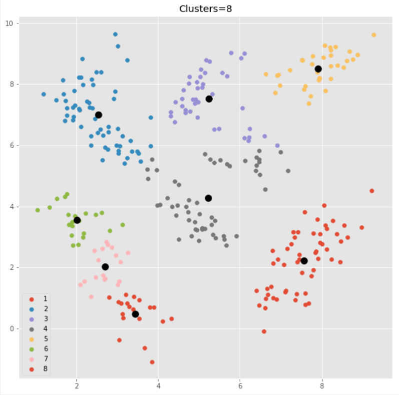
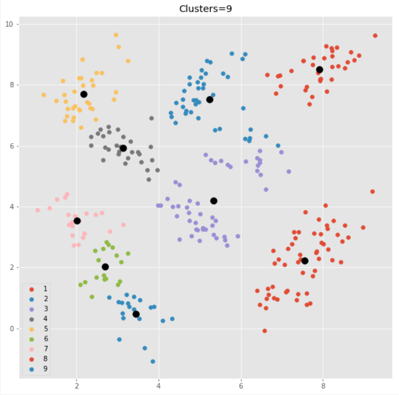
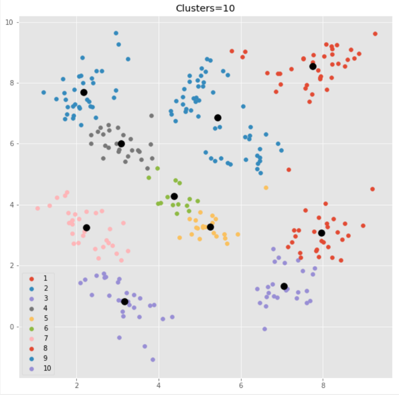


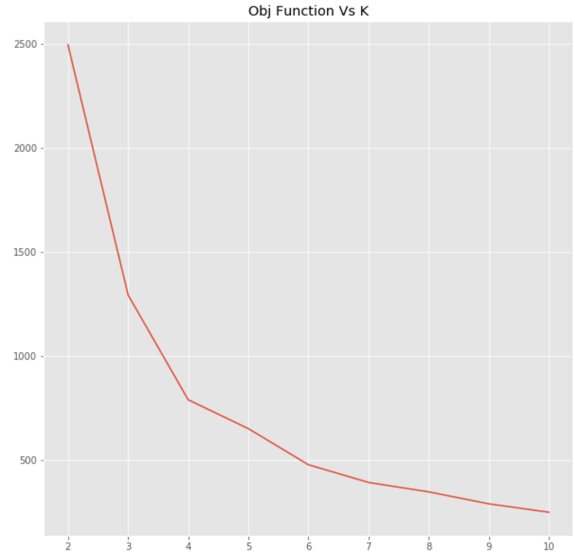
Figure 1- Input Data

K Value	Iniital Centroids	Final Centroids	Objective Func. Value	Clustered Data (black points represent centroids)
K=2	1; 8.36230458, 3.08961725 2; 1.92561853, 2.73857632	1: 6.8071367, 5.40112426 2: 3.01682343, 4.47741928	2498.11	
K=3	1; 6.6384501, 8.33574252 2; 5.07250754, 7.89834048 3; 2.81629029, 3.1999725	1; 6.49724962, 7.52297293 2; 2.56146449, 6.08861338 3; 5.47740039, 2.25498103	1293.77	
K=4	1; 7.30246332 3.16580577 2; 3.79752017 0.69134312 3; 6.2091503 6.16038763 4; 3.9649361 5.20027567	1; 7.25262683 2.40015826 2; 3.24285347 2.55197905 3; 6.57957643 7.57333595 4; 2.8337661 6.9189569	788.964	

K=5	1; 3.2115245, 1.1089788 2; 2.18568667, 3.11739024 3; 5.01728788, 3.76311975 4; 2.10054891, 1.44144019 5; 7.80003043, 1.90963115	1; 3.14506148 0.90770655 2; 2.70510783 6.98765539 3; 6.51196671 7.5619758 4; 3.49556658 3.56611232 5; 7.41419243 2.32169114	650.149	
K=6	1; 4.05095774, 4.05212767 2; 5.27137631, 5.53516715 3; 2.78903847, 6.44350728 4; 2.97097541, 2.39669382 5; 1.05217427, 3.88943741 6; 3.01047612, 6.54286455	1; 7.41419243, 2.32169114 2; 7.75648325, 8.55668928 3; 2.56333815, 6.9782248 4; 3.14506148, 0.90770655 5; 3.49556658, 3.56611232 6; 5.46427736, 6.83771354	476.118	
K=7	1; 7.56399709 7.83135288 2; 4.9511002 8.08344216 3; 1.05217427 3.88943741 4; 1.69565649 7.68082458 5; 4.6733967 7.14753742 6; 7.74867074 1.71812324 7; 3.03696341 5.82211317	1; 7.91430998 8.51990981 2; 5.0217766 7.82401258 3; 2.68198633 2.09461587 4; 2.56333815 6.9782248 5; 6.15468228 5.70140721 6; 7.55616782 2.23516796 7; 4.81833058 3.6950232	390.91	

K=8	1; 3.0226944 0.86402039 2; 2.3085098 7.39324133 3; 4.91688902 7.51334885 4; 5.52279832 5.52162016 5; 8.22144628 8.60551337 6; 2.04945194 2.75937105 7; 2.5366924 2.24222672 8; 4.10720306 0.2505651	1; 3.44650803 0.47784504 2; 2.54165252 7.00267832 3; 5.24028296 7.53131029 4; 5.23053667 4.2793425 5; 7.91430998 8.51990981 6; 2.00857179 3.54850646 7; 2.69805343 2.0242299 8;7.55616782 2.23516796	345.58	
K=9	1; 8.37895231 8.62509614 2; 4.10720306 0.25056515 3; 6.4095594 5.35040201 4; 3.49606966 5.79440796 5; 3.32202131 6.15602339 6; 3.2115245 1.1089788 7; 1.89256383 3.05142539 8; 7.80003043 1.90963115 9; 6.05509889 7.23007608	1; 7.91430998 8.51990981 2; 3.44650803 0.47784504 3; 5.34560332 4.20335478 4; 3.13834768 5.93372322 5; 2.18321462 7.70355341 6; 2.69805343 2.0242299 7; 2.00857179 3.54850646 8; 7.55616782 2.23516796 9; 5.24028296 7.53131029	288.10	
K=10	1;5.14255397 8.37451307 2; 2.48989693 8.40047863 3; 6.39627447 1.24125663 4; 3.12914724 3.40388727 5; 5.33498937 3.07430754 6; 5.2979492 3.65258141 7; 3.08143147 2.18786562 8; 7.57805025 3.82487017 9; 6.47011829 5.54035543 10;3.12073696 0.48979079	1; 7.75648325 8.55668928] 2; 2.18321462 7.70355341] 3; 7.05668293 1.33319679] 4; 3.08507778 5.99914802] 5; 5.25113546 3.2786817] 6; 4.37521312 4.27426189] 7; 2.24204752 3.25100749] 8; 7.95957401 3.08441042] 9; 5.43207068 6.86930884] 10;3.16906145 0.81432515	247.55	

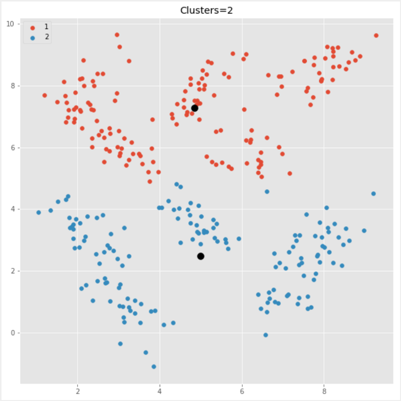
And the objective function for different K-values(strategy 1) are plotted as shown below

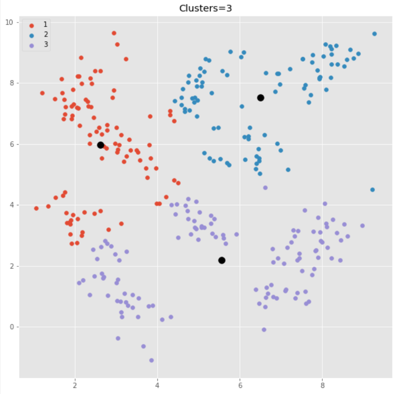
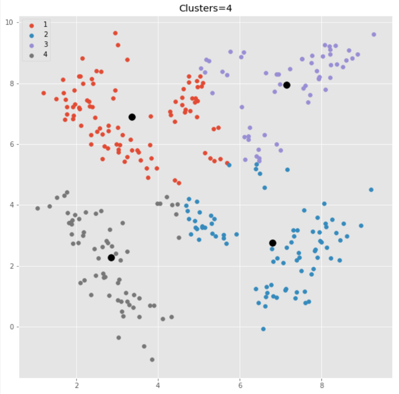
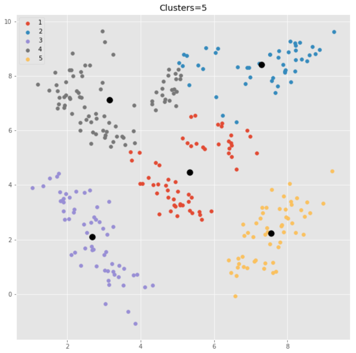
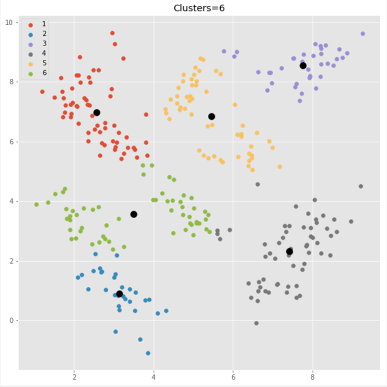


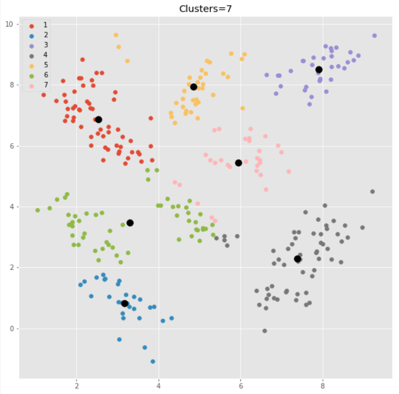
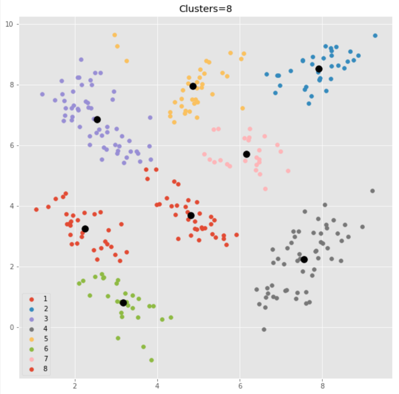
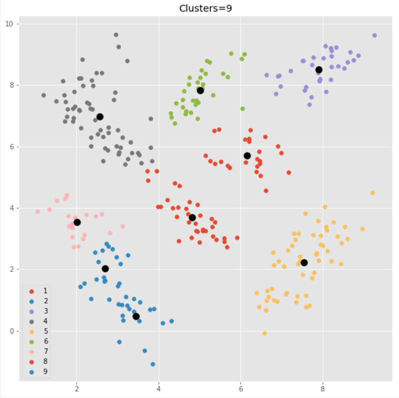
Strategy II:

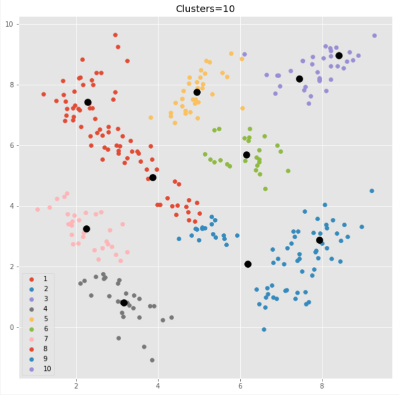
The only difference between strategy 1 and strategegu 2 is the initalisation of the centroids. In strategy 2 first a centroid is chosen randomly, from that centroid disance to each sample is calculated and the sample which has a maximal average distance to the previous centroid is taken as the new centroid. Now as two centroids are obtained for the third centroid, again ditances from each previous centroid to all the samples are calculated then the average distance is computed and the sample which has the maximal average distance is taken as a new centroid. The above steps are repeated for centroids from 2-10. Now as the inital centroids are obtained the rest of the process is the same as strategy 1.

Results for Strategy II:

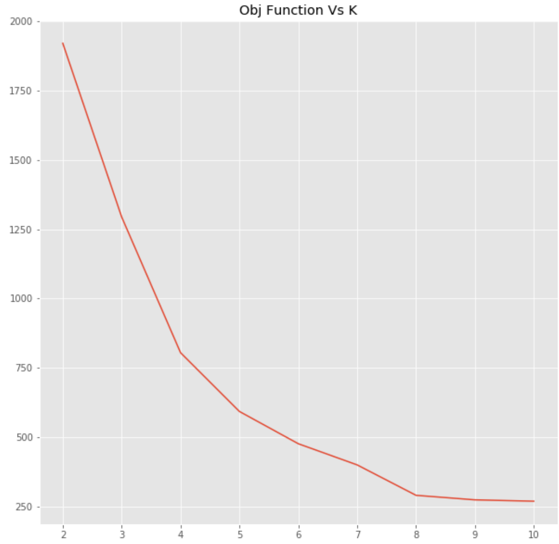
K Value	Initial Centroids	FInal Centroids	Objective Func. Value	Clustered Data (black points represent centroids)
K=2	1; 7.22537424 8.46609363 2; 3.85212146 -1.08715226	1; 4.85261193 7.27164171 2; 5.00056234 2.48542748	1921.03	

K=3	1; 2.73285832 2.83024707 2; 9.26998864 9.62492869 3; 3.85212146 -1.08715226	1; 2.61946868 5.96519477 2; 6.49724962 7.52297293 3; 5.55524182 2.18980958	1294.29	
K=4	1; 2.05924902 7.20598798 2; 6.5807212 -0.0766824 3; 9.26998864 9.62492869 4; 3.85212146 -1.08715226	1; 3.36759466 6.90961066 2; 6.80866964 2.75651994 3; 7.14834495 7.96153683 4; 2.85235149 2.28186483	804.65	
K=5	1; 3.72610844 5.20432439 2; 9.26998864 9.62492869 3; 3.85212146 -1.08715226 4; 2.95297924 9.65073899 5; 6.5807212 -0.0766824	1; 5.33907212 4.46551175 2; 7.29974969 8.41331838 3; 2.68198633 2.09461587 4; 3.15072761 7.12192906 5; 7.55616782 2.23516796	592.528	
K=6	1; 2.46087695 6.86898874 2; 3.85212146 -1.08715226 3; 9.26998864 9.62492869 4; 7.68097556 0.83542043 5; 2.95297924 9.65073899 6; 3.04101702 -0.36138487	1; 2.56333815 6.9782248 2; 3.14506148 0.90770655 3; 7.75648325 8.55668928 4; 7.41419243 2.32169114 5; 5.46427736 6.83771354 6; 3.49556658 3.56611232	476.118	

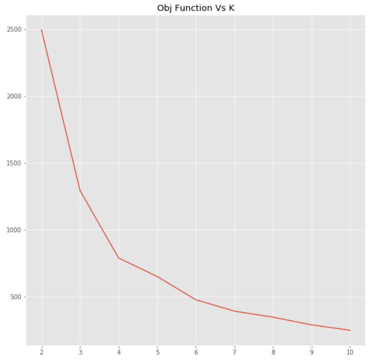
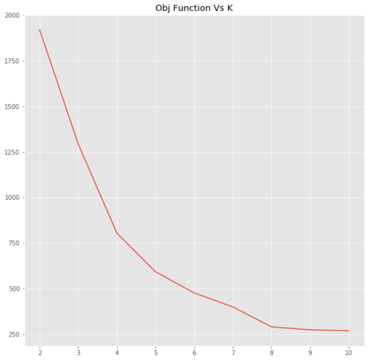
K=7	1; 2.38952606 7.22195564 2; 3.85212146 -1.08715226 3; 9.26998864 9.62492869 4; 7.68097556 0.83542043 5; 2.95297924 9.65073899 6; 3.04101702 -0.36138487 7; 8.87578072 8.96092361	1; 2.53650108 6.85941978 2; 3.16906145 0.81432515 3; 7.91430998 8.51990981 4; 7.39380325 2.29452245 5; 4.85939875 7.94163821 6; 3.31074837 3.47473078 7; 5.94696208 5.44598487	399.68	
K=8	1; 3.66118224 -0.63372377 2; 9.26998864 9.62492869 3; 1.20162248 7.68639714 4; 7.68097556 0.83542043 5; 2.95297924 9.65073899 6; 3.85212146 -1.08715226 7; 8.87578072 8.96092361 8; 3.04101702 -0.36138487	1; 2.24204752 3.25100749 2; 7.91430998 8.51990981 3; 2.53650108 6.85941978 4; 7.55616782 2.23516796 5; 4.85939875 7.94163821 6; 3.16906145 0.81432515 7; 6.15468228 5.70140721 8; 4.81833058 3.6950232	289.9	
K=9	1; 6.60277235 6.31081582 2; 3.85212146 -1.08715226 3; 9.26998864 9.62492869 4; 1.20162248 7.68639714 5; 6.5807212 -0.0766824 6; 2.95297924 9.65073899 7; 3.04101702 -0.36138487 8; 8.87578072 8.96092361 9; 3.66118224 -0.63372377	1; 4.81833058 3.6950232 2; 3.44650803 0.47784504 3; 7.91430998 8.51990981 4; 2.56333815 6.9782248 5; 7.55616782 2.23516796 6; 5.0217766 7.82401258 7; 2.00857179 3.54850646 8; 6.15468228 5.70140721 9; 2.69805343 2.0242299	273.57	

K=10	1; 2.64683045 6.32344268 2; 6.5807212 -0.0766824 3; 9.26998864 9.62492869 4; 3.85212146 -1.08715226 5; 2.95297924 9.65073899 6; 8.87578072 8.96092361 7; 3.04101702 -0.36138487 8; 1.20162248 7.68639714 9; 7.68097556 0.83542043 10; 8.678057 9.08757916	1; 3.8596884 4.94757973 2; 6.1829665 2.0830502 3; 8.41127011 8.97490383 4; 3.16906145 0.81432515 5; 4.95254423 7.76039378 6; 6.15468228 5.70140721 7; 2.24204752 3.25100749 8; 2.28840393 7.42784851 9; 7.94171396 2.87966135 10;7.45085073 8.20356187	268.72	

And the objective function (strategy 2) for different K-values are plotted as shown below



After running the code number of times with different initalizations and comparing the objective functions for both the strategies, the optimal clusters is 4. The results obtained are just for one random initalization.

Strategy 1	Strategy 2
 Obj Function Vs K 2500 2000 1500 1000 500 2 3 4 5 6 7 8 9 10	 Obj Function Vs K 2000 1750 1500 1250 1000 750 500 250 2 3 4 5 6 7 8 9 10