Assignment Solution

Answer 1-

In order to identify the costumers based on the available data the selection criteria should be based on below primary key factors :

1-Which supplier has high gross $value of contract

2-Who is having more number of contract

3-Who is having more duration of contract

Approach :

To meet our goal we will follow the elimination method. After analyzing the dataset on the basis of primary key factors we have extracted important features or fields which are :

1-**Agency_Name** – Name of agency responsible for contract work

2-**Start_Date** – Date from which contracted work is going to commence.

3-**End_Date** – Day at which contracted work will finish.

4-**Contract_days** – Number of working days calculated by difference between End date and start data.

5-**Value** – Amount which agency required from bank for contract.

Based on the number of contracts below is the list of top 5 agencies in descending order :

Agency Name	Number of contracts
Department of Defence	26850
Department of Veterans' Affairs	3663
Services Australia	3082
Australian Taxation Office	2841
Department of Home Affairs	2654

Based on total gross value below is the list of top 5 agencies in descending order :

Agency_Name	Value
Independent Parliamentary Expenses Authority	974959.1
Australian Taxation Office	9708162000
Great Barrier Reef Marine Park Authority	93808460
Department of Agriculture, Water and the Environment	900538000
Department of Industry, Science, Energy and Resources	842642000

Based on total contracted days below is the list of top 5 agencies in descending order :

Agency_Name	Contract days
Department of Defence	13005485
Department of Home Affairs	1973775
Australian Taxation Office	1258733
Department of Veterans' Affairs	1023941
Australian Federal Police	918947

Another important selection criteria factor could be Pay rate of agencies which can be calculated by using formula :

Pay rate = (Total gross value) / (Number of contracted days)

Based on total Pay rate below is the list of top 10 agencies in descending order :

Agency_Name	Gross value	Contract_days	Rate_per_day
Digital Transformation Agency	1530997000	7300	209725.5547
Department of Defence	3.38203E+11	13005485	26004.63648
Regional Investment Corporation	21524910	1534	14031.88349
Australian Pesticides and Veterinary Medicines Authority	35928550	3285	10937.15288
Sydney Harbour Federation Trust	18091970	1847	9795.328587
Department of Health	18904780	2034	9294.384105
Department of Foreign Affairs and Trade	1533824000	173118	8859.988606
Australian Taxation Office	9708162000	1258733	7712.645673
Department of Home Affairs	14470770000	1973775	7331.519353
Department of Finance	2700382000	569203	4744.1457

If we consider all the factors and based on the above evidences we can conclude that agencies **Department of Defence, Australian Taxation Office, Department of Home Affairs** falls consistently in the list of high Gross value, number of contract days and Rate per day thus Bank X should consider providing services to them.

Answer 2-

In order to identify the new office locations based on the available data the selection criteria should be based on below primary key factors :

1- Which pincode has maximum occurrence

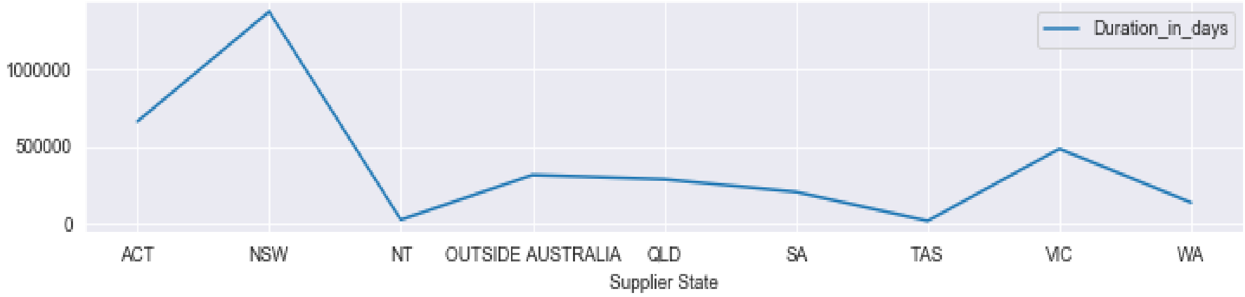
2- Plot the map or identify clusters based on pin code and place the office based on proximities to reduce the cost since it is more in cities.

Approach :

To meet our goal we will follow the elimination method. After analyzing the dataset on the basis of primary key factors we have extracted important features or fields which are Supplier State and number of contract days.

Below is the list and visualization of Number of contracted days with respect to states so that we can identify which states have more work.

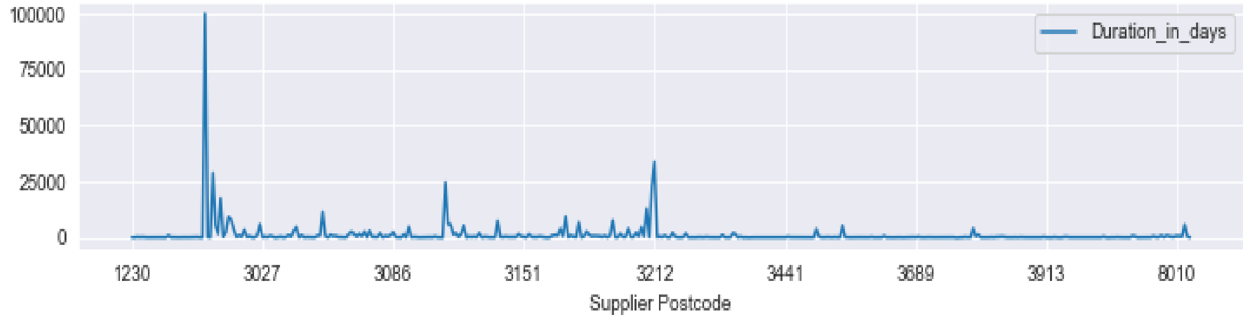
Supplier State	Contract days
ACT	659450
NSW	1369550
NT	24183
OUTSIDE AUSTRALIA	313249
QLD	285618
SA	203646
TAS	16767
VIC	482607
WA	135471



Based on the above graphical illustration and tabular data we can say that we have 4 significant states which are ACT, NSW,QLD and VIC and our further analysis will be based on these 4 states.

VIC state analysis :

Based on the number of contract days according to pincode below is the graphical illustration :

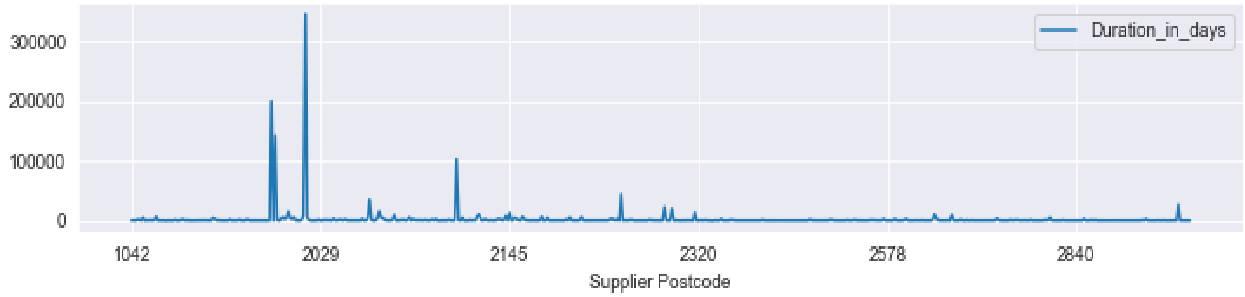


As per above graph it is evident that there is a spike near pincode 3000, 3100 and 3212 which indicates significant presence of suppliers in these regional clusters.

Preferred office location : Based on the above graphical representation pincode 3170 which is for region Mulgrave seems to be the most suitable one as it have comparatively cheaper rent and can be accessed easily by all the major suppliers.

NSW state analysis :

Based on the number of contract days according to pincode below is the graphical illustration :

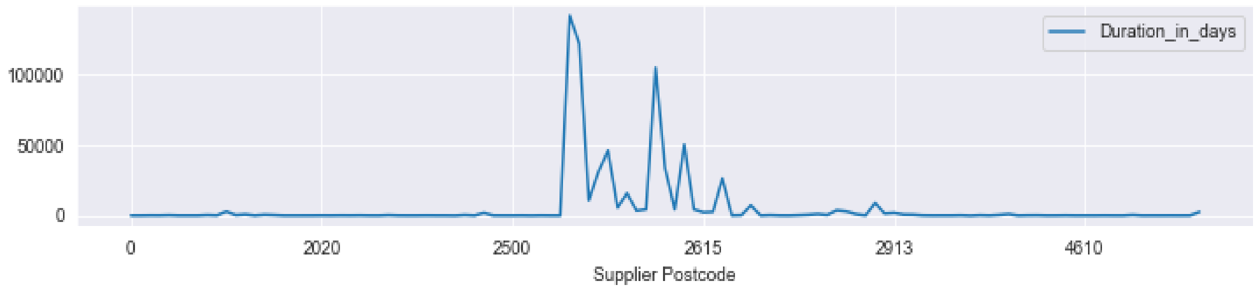


As per above graph it is evident that there is a spike near pincode 2000, 2100 and 2300 which indicates significant presence of suppliers in these regional clusters.

Preferred office location : Based on the above graphical representation pincode 2113 which is for region Macquarie Park seems to be the most suitable one as it have comparatively cheaper rent and can be accessed easily by all the major suppliers.

ACT state analysis :

Based on the number of contract days according to pincode below is the graphical illustration :

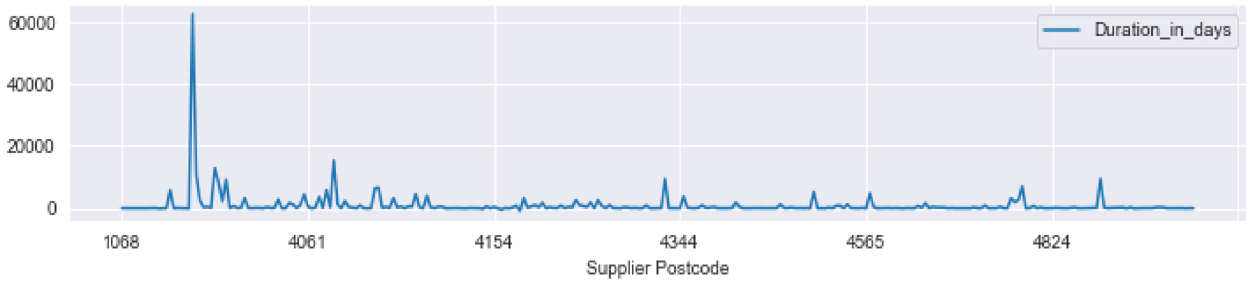


As per above graph it is evident that there is a spike between pincode 2550 and 2615 which indicates significant presence of suppliers in these regional clusters.

Preferred office location : Based on the above graphical representation pincode 2617 which is for region Mckellar seems to be the most suitable one as it have comparatively cheaper rent and can be accessed easily by all the major suppliers.

QLD state analysis :

Based on the number of contract days according to pincode below is the graphical illustration :



As per above graph it is evident that there is a spike near pincode 1550 and 4060 which indicates significant presence of suppliers in these regional clusters.

Preferred office location : Based on the above graphical representation pincode 4020 which is for region Redcliffe seems to be the most suitable one as it have comparatively cheaper rent and can be accessed easily by all the major suppliers.

As per above discussion based on Suppliers postcode preferred new office locations are as below :

Place	Pin code	Country	State
Mckellar	2617	Australia	ACT
Macquarie Park	2113	Australia	NSW
Redcliffe	4020	Australia	QLD
Mulgrave	3170	Australia	VIC

Answer 3 :

Problem statement : Identify new office locations for Bank X to serve suppliers representative better.

Proposed solution : Our idea of solution towards identifying new office locations revolves around the concept of suppliers pincode which has maximum occurrence or frequency in dataset and try to identify clusters based on suppliers pin code, suppliers contract days and place the office based on proximities to reduce the cost since it is more in cities.

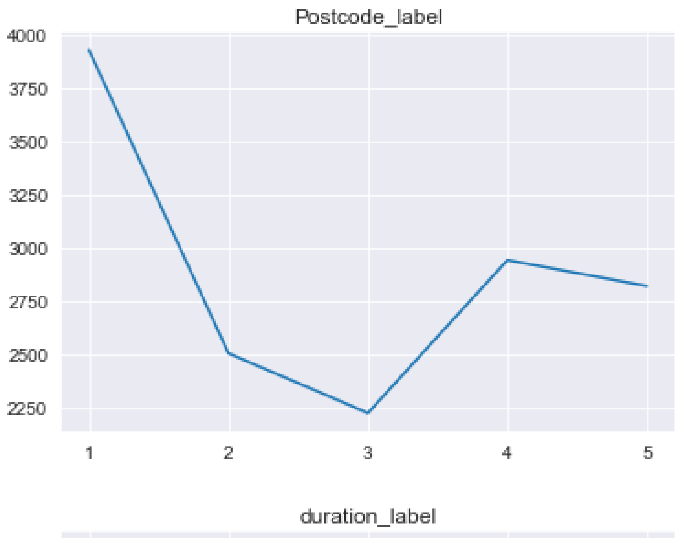
Machine learning algorithm and implementation – Our proposed solution is based on suppliers pincode segmentation so the solution comes under the category of unsupervised learning thus **K-means clustering** is the best suited one for this problem.

We are implementing sample machine learning specifically for Victoria state.

We have considered Supplier Postcode and contract days as key features for segmentation or clustering process. Since we have lots of unique suppliers postcode, in order to get more insight we have applied frequency binning technique on postcode according and their respective contract days to identify which clusters are more significant. We have divided above mentioned features into 5 segments using frequency binning technique.

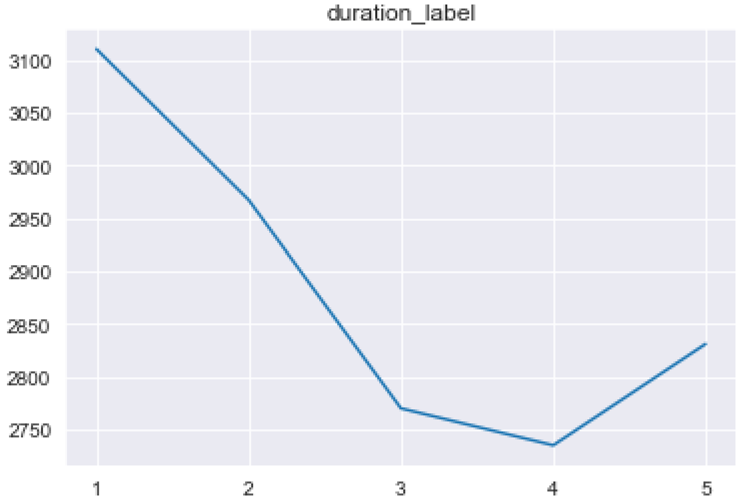
Below is the postcode segmentation and graphical illustration as per cluster :

Postcode range	Total Freequency	Cluster
(1229.999, 3000.0]	3929	1
(3109.0, 3180.0]	2941	2
(3180.0, 30004.0]	2819	3
(3000.0, 3008.0]	2502	4
(3008.0, 3109.0]	2222	5



Below is the contract days segmentation and graphical illustration as per cluster :

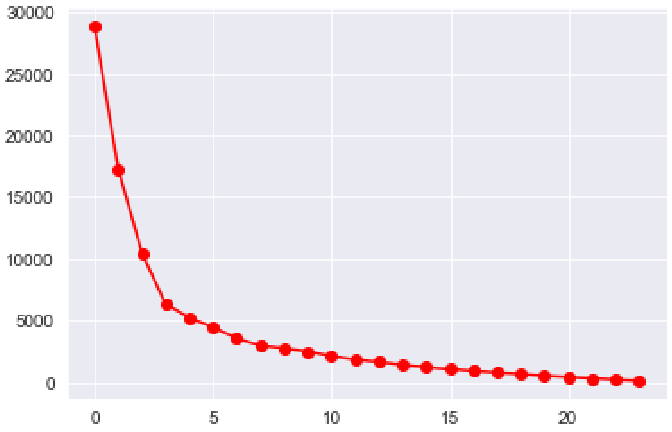
Contract days range	Total Contract days	Cluster
(-475.001, 1.0]	3110	1
(1.0, 7.0]	2967	2
(28.0, 6575.0]	2831	3
(7.0, 14.0]	2770	4
(14.0, 28.0]	2735	5



It is evident from above graphical illustrations that comparatively cluster 1 and 4 postcodes have high frequency and cluster 1,2 postcodes have high contract days.

Moving ahead we normalize our dataset in range of 0-1 and apply **Elbow Method** to choose an optimal value of K which will be used further for clustering.

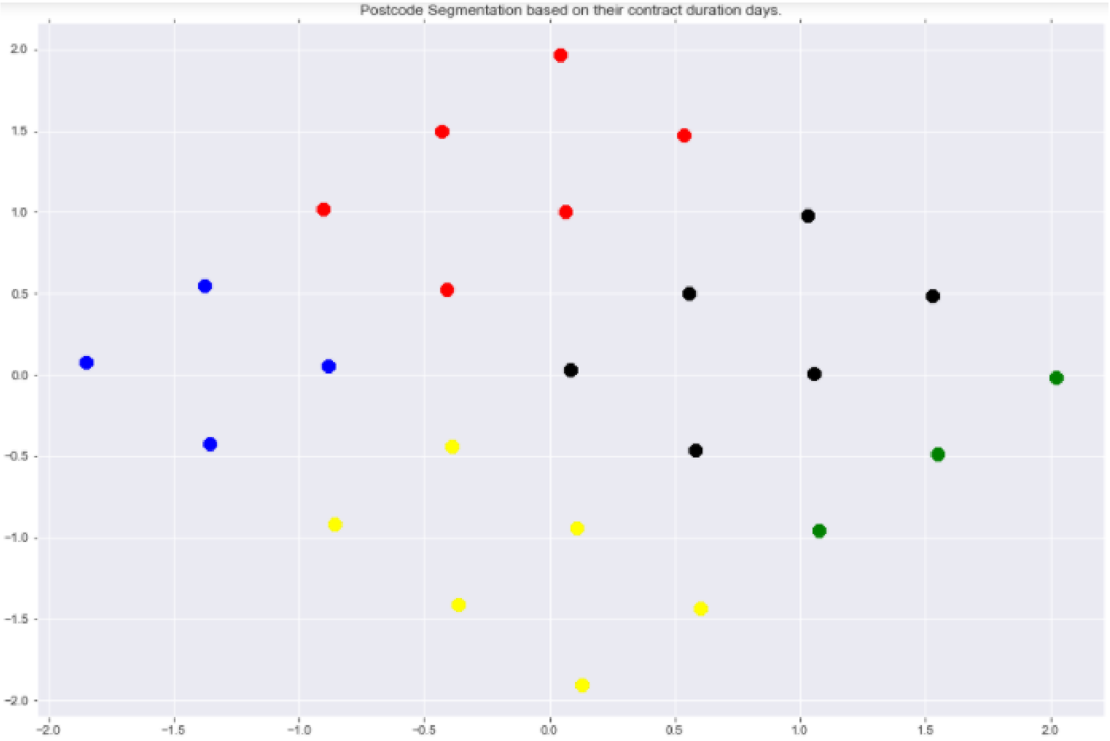
Below is the graph for each of the K values where we calculate average distances to the centroid across all data points.



We try to find a point where the average distance from the centroid falls suddenly and that point is K=5.

We apply concept of Principal Component Analysis (PCA) to make sure that we capture as many features of the original dataset as possible which is required for plotting of clusters.

Below is the scatter plot of Post code segmentation based on their respective number of contract days.



Color	Cluster number
Red	0
Blue	1
Green	2
Yellow	3
Black	4

Pros –

1- Simple : It was easy to implement in order to identify unknown groups or clusters of data.

2- Suitable in a large dataset – Our dataset is large and K-means is suitable for a large number of datasets and it's computed much faster.

3- Easy to interpret - k-means produce tighter clusters and results are easy to interpret.

Cons-

1- Uniform effect - It produces cluster with uniform size even when the input data has different sizes which can be seen in above graph

2- Sensitivity to scale- We have applied normalization and changing or rescaling the dataset either through normalization or standardization may completely change the final results.

3- Handle numerical data- K-means algorithm can be performed only in numerical data which restricts our scope of including more features.

References :

1- Analytics Vidhya
2- Towardsdatascience
3- Rmit university lecture slides

Author : Anirudhda Pardhi

Student id : s3807109

Mail id : Anirudhda.pardhi@gmail.com