
CSI – 695 Final Project

PRESENTED BY: AJAY KULKARNI (Go1024139)

Outline

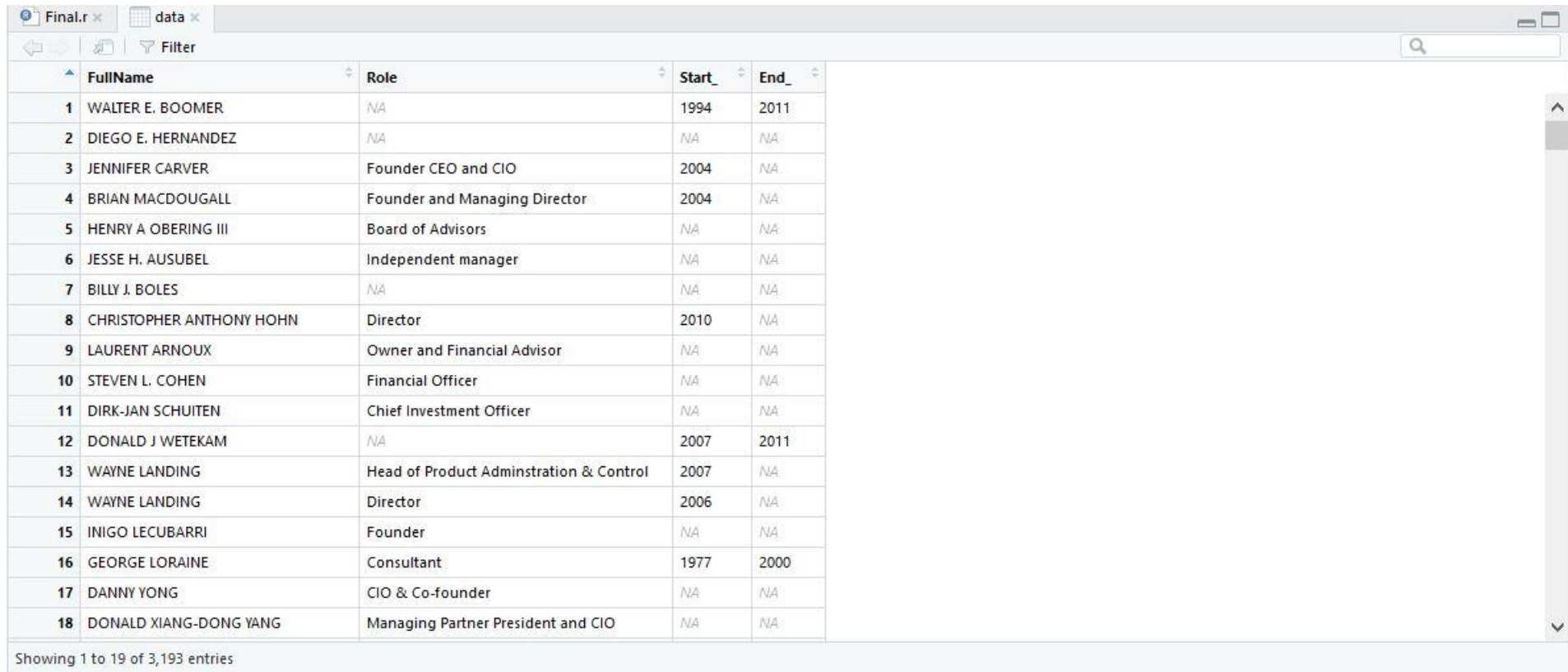
- Data import from MySQL to R using “RMySQL”.
- Data cleaning and processing.
- Results of kNN and kd-tree.
- Comparison of kNN and kd-tree based on execution time.

Data import

- The important attributes considered for finding similarity are
 - Role
 - Start Year
 - End Year
- The data is imported into R using the following query

```
“SELECT people.FullName, peopleandcompanies.Role, peopleandcompanies.Start_,  
peopleandcompanies.End_ FROM people INNER JOIN peopleandcompanies ON  
people.Dir_ID=peopleandcompanies.People_ID”
```

Data import



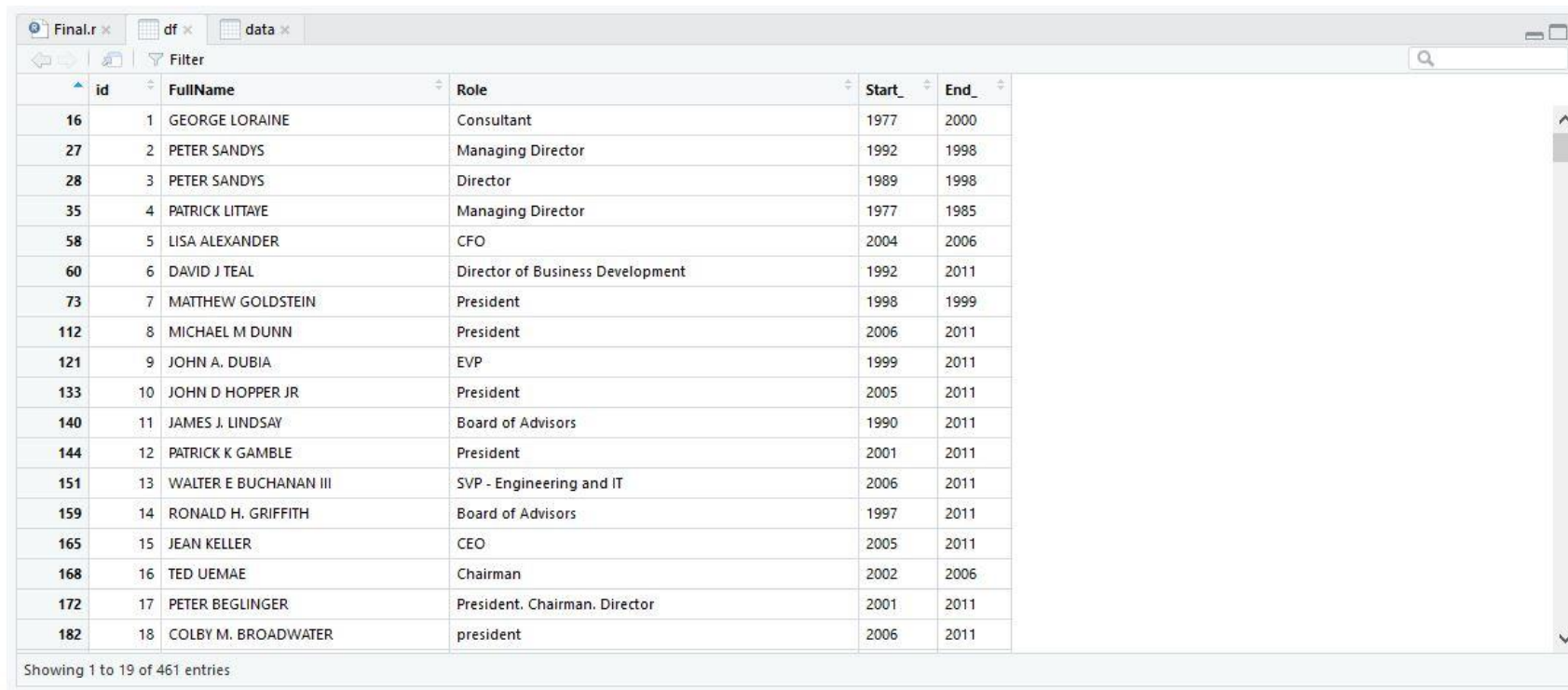
The screenshot shows an RStudio window with a data frame named 'data' loaded. The data frame has 19 columns and 3,193 rows. The first 18 rows are visible, showing names, roles, and dates. The columns are: FullName, Role, Start_, and End_. The data is sorted by Start_ date.

	FullName	Role	Start_	End_
1	WALTER E. BOOMER	NA	1994	2011
2	DIEGO E. HERNANDEZ	NA	NA	NA
3	JENNIFER CARVER	Founder CEO and CIO	2004	NA
4	BRIAN MACDOUGALL	Founder and Managing Director	2004	NA
5	HENRY A OBERING III	Board of Advisors	NA	NA
6	JESSE H. AUSUBEL	Independent manager	NA	NA
7	BILLY J. BOLES	NA	NA	NA
8	CHRISTOPHER ANTHONY HOHN	Director	2010	NA
9	LAURENT ARNOUX	Owner and Financial Advisor	NA	NA
10	STEVEN L. COHEN	Financial Officer	NA	NA
11	DIRK-JAN SCHUITEN	Chief Investment Officer	NA	NA
12	DONALD J WETEKAM	NA	2007	2011
13	WAYNE LANDING	Head of Product Adminstration & Control	2007	NA
14	WAYNE LANDING	Director	2006	NA
15	INIGO LECUBARRI	Founder	NA	NA
16	GEORGE LORRAINE	Consultant	1977	2000
17	DANNY YONG	CIO & Co-founder	NA	NA
18	DONALD XIANG-DONG YANG	Managing Partner President and CIO	NA	NA

Showing 1 to 19 of 3,193 entries

Data cleaning and processing

- NAs in the data are removed using `na.omit()` function in R and “id” column has been added as an identification for each data point.



	id	FullName	Role	Start_	End_
16	1	GEORGE LORAINÉ	Consultant	1977	2000
27	2	PETER SANDYS	Managing Director	1992	1998
28	3	PETER SANDYS	Director	1989	1998
35	4	PATRICK LITTAYE	Managing Director	1977	1985
58	5	LISA ALEXANDER	CFO	2004	2006
60	6	DAVID J TEAL	Director of Business Development	1992	2011
73	7	MATTHEW GOLDSTEIN	President	1998	1999
112	8	MICHAEL M DUNN	President	2006	2011
121	9	JOHN A. DUBIA	EVP	1999	2011
133	10	JOHN D HOPPER JR	President	2005	2011
140	11	JAMES J. LINDSAY	Board of Advisors	1990	2011
144	12	PATRICK K GAMBLE	President	2001	2011
151	13	WALTER E BUCHANAN III	SVP - Engineering and IT	2006	2011
159	14	RONALD H. GRIFFITH	Board of Advisors	1997	2011
165	15	JEAN KELLER	CEO	2005	2011
168	16	TED UEMAE	Chairman	2002	2006
172	17	PETER BEGLINGER	President, Chairman, Director	2001	2011
182	18	COLBY M. BROADWATER	president	2006	2011

Showing 1 to 19 of 461 entries

Data cleaning and processing

- After performing data cleaning data contains 461 rows and 5 columns.
- One hot encoding has been performed on all the selected attributes.

[illegible]

Results of kNN and kd-tree

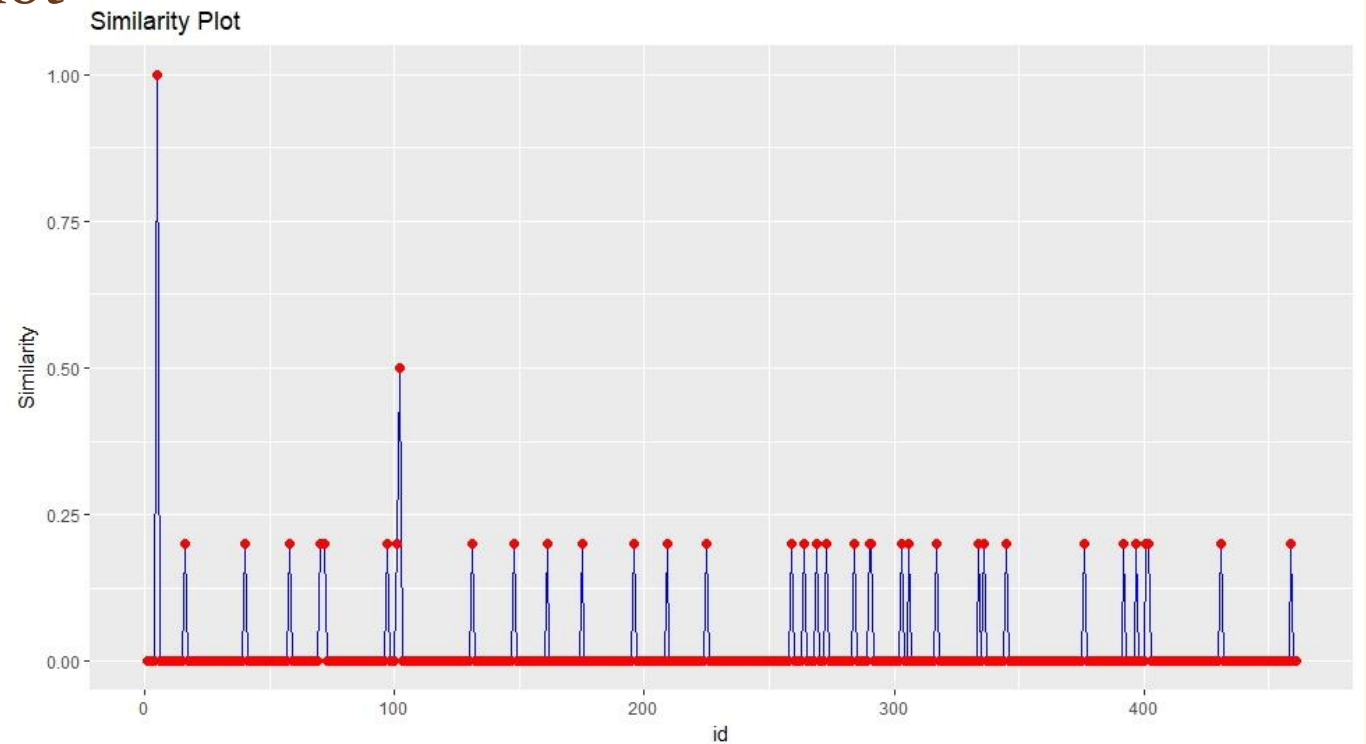
- For kNN, the binary similarity is used to find the similarity between different objects.
- KNN and kd-tree are tested on following examples.

Id	K
5	1
30	5
100	10
243	20
343	50

Example – 1 (id = 5, k =1)

- KNN result and similarity plot

```
Console  Terminal x
~/
> myknn(5,1)
[1] "Execution time"
8.69 sec elapsed
[1] "Binary Similarity"
[1] 5 102
[1] "LISA ALEXANDER" "RICHARD BATTEY"
[1] "CFO" "CFO"
[1] 1.0 0.5
> |
```



Example – 1 (id = 5, k =1)

KNN

```
Console Terminal x
~/
> myknn(5,1)
[1] "Execution time"
8.69 sec elapsed
[1] "Binary similarity"
[1] 5 102
[1] "LISA ALEXANDER" "RICHARD BATTEY"
[1] "CFO" "CFO"
[1] 1.0 0.5
> |
```

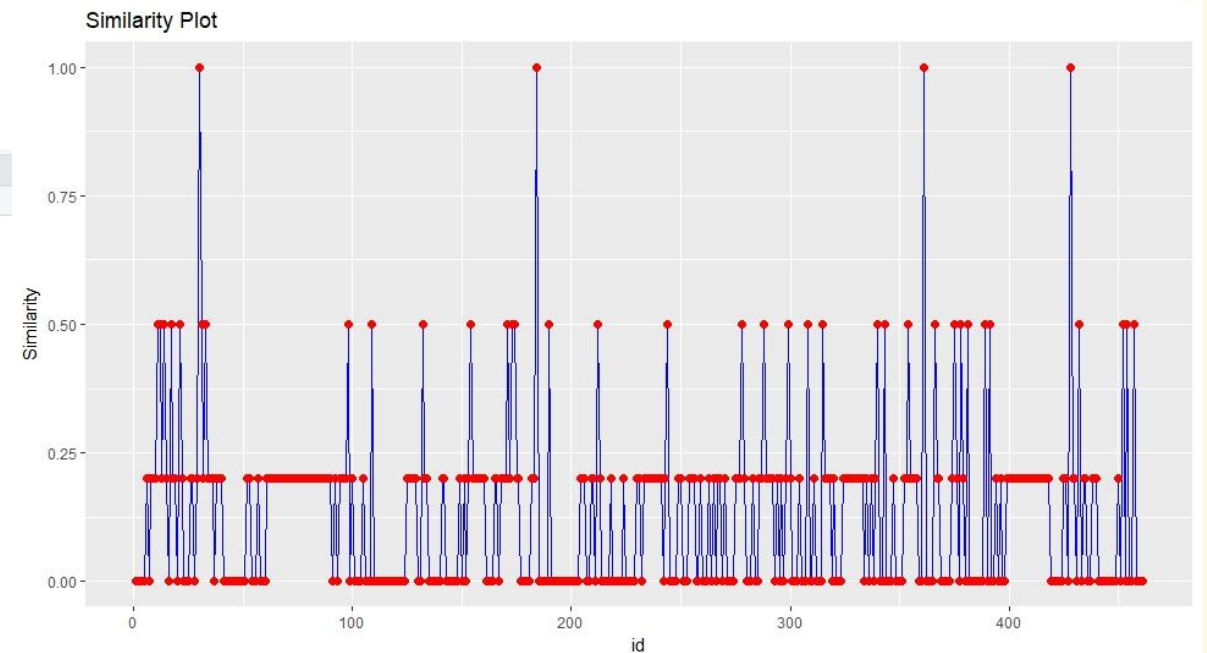
Kd-tree

```
Console Terminal x
C:/Users/srishti/Desktop/ajay/
> mykdtree(5,1)
[1] "kd-tree results"
[1] "Execution Time"
0.12 sec elapsed
[1] 5 102
[1] "LISA ALEXANDER" "RICHARD BATTEY"
[1] "CFO" "CFO"
[1] 0 2
> |
```

Example – 2 (id = 30, k = 5)

- KNN result and similarity plot

```
Console Terminal x
C:/Users/srishiti/Desktop/ajay/
> myknn(30,5)
[1] "Execution time"
8.66 sec elapsed
[1] "Binary Similarity"
[1] 30 184 361 428 11 12
[1] "FREDERICK MCCORKLE" "CHARLES T ROBERTSON" "HENRY H. SHELTON" "PAUL DAVID MILLER"
[5] "JAMES J. LINDSAY" "PATRICK K GAMBLE"
[1] "Board of Advisors" "Board of Advisors" "Board of Advisors" "Board of Advisors" "Board of Advisors"
[6] "President"
[1] 1.0 1.0 1.0 1.0 0.5 0.5
>
```



Example – 2 (id = 30, k =5)

KNN

```
Console Terminal x
C:/Users/srishiti/Desktop/ajay/
> myknn(30,5)
[1] "Execution time"
8.66 sec elapsed
[1] "Binary Similarity"
[1] 30 184 361 428 11 12
[1] "FREDERICK MCCORKLE" "CHARLES T ROBERTSON" "HENRY H. SHELTON" "PAUL DAVID MILLER"
[5] "JAMES J. LINDSAY" "PATRICK K GAMBLE"
[1] "Board of Advisors" "Board of Advisors" "Board of Advisors" "Board of Advisors" "Board of Advisors"
[6] "President"
[1] 1.0 1.0 1.0 1.0 0.5 0.5
>
```

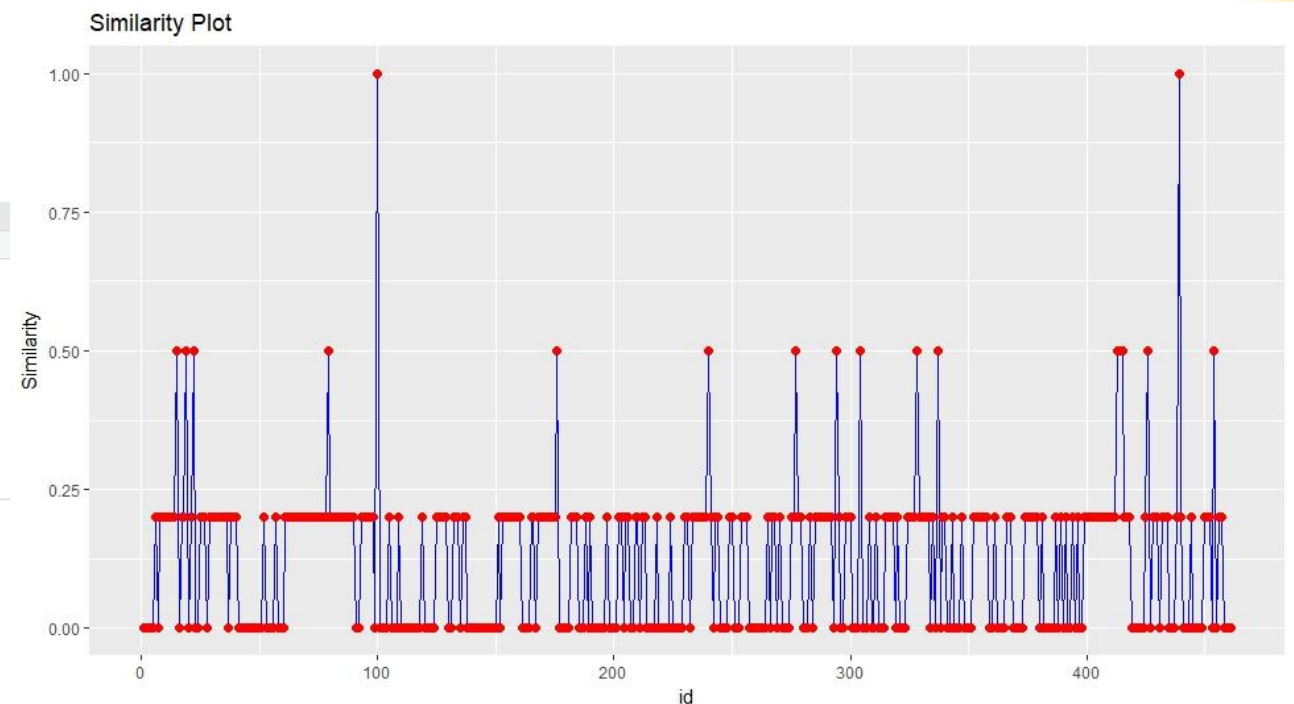
Kd-tree

```
Console Terminal x
C:/Users/srishiti/Desktop/ajay/
> mykdtree(30,5)
[1] "kd-tree results"
[1] "Execution Time"
0.13 sec elapsed
[1] 30 428 361 184 173 381
[1] "FREDERICK MCCORKLE" "PAUL DAVID MILLER" "HENRY H. SHELTON" "CHARLES T ROBERTSON"
[5] "ARTHUR J LICHTE" "HENRY A OBERING III"
[1] "Board of Advisors" "Board of Advisors" "Board of Advisors" "Board of Advisors" "Board of Advisors"
[6] "Board of Advisors"
[1] 0 0 0 0 2 2
>
```

Example – 3 (id = 100, k = 10)

- KNN result and similarity plot

```
Console Terminal x
C:/Users/srishti/Desktop/ajay/
> myknn(100,10)
[1] "Execution time"
8.44 sec elapsed
[1] "Binary similarity"
[1] 100 439 15 19 22 79 176 240 277 294 304
[1] "EDWIN H. BURBA" "MATTHEW COOPER" "JEAN KELLER" "EDWIN S. LELAND"
[5] "EDGAR R. ANDERSON JR." "WILLIAM HARTZOG" "CHARLES PITMAN" "JEROME B. HILMES"
[9] "JIMMIE V. ADAMS" "JOHN M. NOWAK" "GEORGE CHRISTMAS"
[1] "CEO" "CEO" "CEO" "President" "CEO" "CEO" "CEO" "CEO"
[9] "Consultant" "CEO" "CEO"
[1] 1.0 1.0 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5
>
```



Example – 3 (id = 100, k =10)

KNN

```

Console Terminal x
C:/Users/srishti/Desktop/ajay/
> myknn(100,10)
[1] "Execution time"
8.44 sec elapsed
[1] "Binary similarity"
[1] 100 439 15 19 22 79 176 240 277 294 304
[1] "EDWIN H. BURBA" "MATTHEW COOPER" "JEAN KELLER" "EDWIN S. LELAND"
[5] "EDGAR R. ANDERSON JR." "WILLIAM HARTZOG" "CHARLES PITMAN" "JEROME B. HILMES"
[9] "JIMMIE V. ADAMS" "JOHN M. NOWAK" "GEORGE CHRISTMAS"
[1] "CEO" "CEO" "CEO" "President" "CEO" "CEO" "CEO" "CEO"
[9] "Consultant" "CEO" "CEO"
[1] 1.0 1.0 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5
>

```

Kd-tree

[illegible]

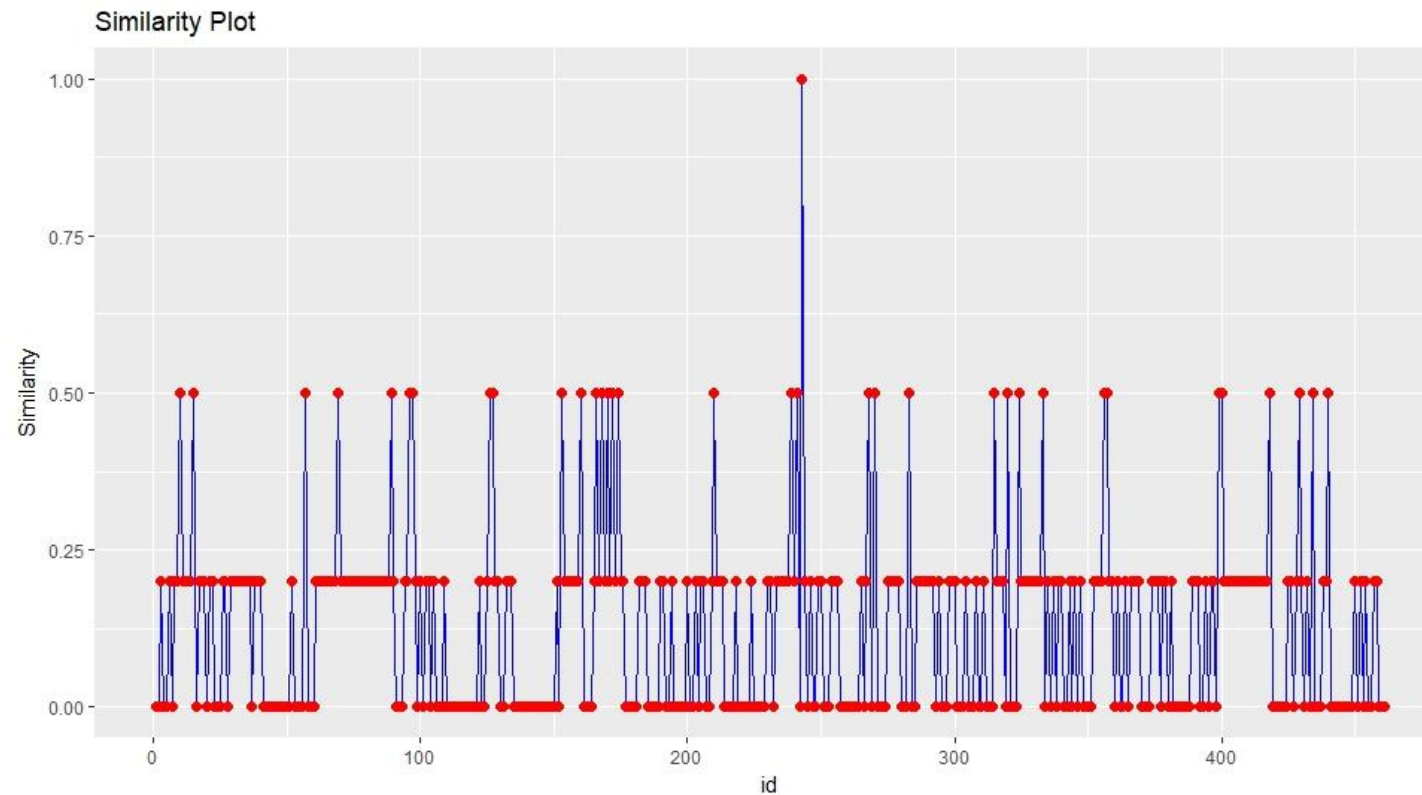
Example - 4 (id = 243, k = 20)

- KNN result

[illegible]

Example – 4 (id = 243, k = 20)

- KNN similarity plot



Example – 4 (id = 243, k =20)

KNN

```
Console Terminal x
C:/Users/srishti/Desktop/ajay/
> myknn(243,20)
[1] "Execution time"
8.23 sec elapsed
[1] "Binary Similarity"
[1] 243 10 15 57 69 89 96 97 126 127 153 160 166 168 170 172 174 210 239 241 268
[1] "WILLIAM L. NYLAND" "JOHN D HOPPER JR" "JEAN KELLER" "CARL G. O'BERRY"
[5] "JOHN P. JUMPER" "DONALD G COOK" "WILLIAM S. WALLACE" "GREGORY G. JOHNSON"
[9] "HARRY RADUEGE" "PAUL J. KERN" "JOHN H. TILELLI" "DENNIS J. REIMER"
[13] "MARVIN D. BRAILSFORD" "PHILIP M. BALISLE" "CHARLES MCCAUSLAND" "BARRY R. MCCAFFREY"
[17] "MICHAEL A HOUGH" "CARL E. MUNDY" "JOHN J. GROSSENBACHER" "RICHARD D. HEARNEY"
[21] "RICHARD A. HACK"
[1] "Director" "President" "CEO" "Director"
[5] "Associate" "Board of Directors" "Director" "Director"
[9] "Senior Counselor" "Senior Counselor" "COO" "Director"
[13] "Director" "vp" "Director" "Director"
[17] "Board of Advisors" "Director" "Director" "Director"
[21] "SVP - Operations"
[1] 1.0 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5
> |
```

Kd-tree

```
Console Terminal x
C:/Users/srishti/Desktop/ajay/
> mykdtree(243,20)
[1] "kd-tree results"
[1] "Execution Time"
0.14 sec elapsed
[1] 243 333 96 283 324 97 239 270 160 241 172 210 57 320 170 166 168 268 357 126 127
[1] "WILLIAM L. NYLAND" "BRUCE CARLSON" "WILLIAM S. WALLACE" "MARTIN PAUL TOLCHER"
[5] "GERALD L. HOEWING" "GREGORY G. JOHNSON" "JOHN J. GROSSENBACHER" "MAXWELL C BAILEY"
[9] "DENNIS J. REIMER" "RICHARD D. HEARNEY" "BARRY R. MCCAFFREY" "CARL E. MUNDY"
[13] "CARL G. O'BERRY" "JOHN R. DAILEY" "CHARLES MCCAUSLAND" "MARVIN D. BRAILSFORD"
[17] "PHILIP M. BALISLE" "RICHARD A. HACK" "BRIAN A ARNOLD" "HARRY RADUEGE"
[21] "PAUL J. KERN"
[1] "Director" "Director" "Director" "Director" "Director"
[6] "Director" "Director" "Director" "Director" "Director"
[11] "Director" "Director" "Director" "Director" "Director"
[16] "Director" "vp" "SVP - operations" "Space Systems" "Senior Counselor"
[21] "Senior Counselor"
[1] 0 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
> |
```

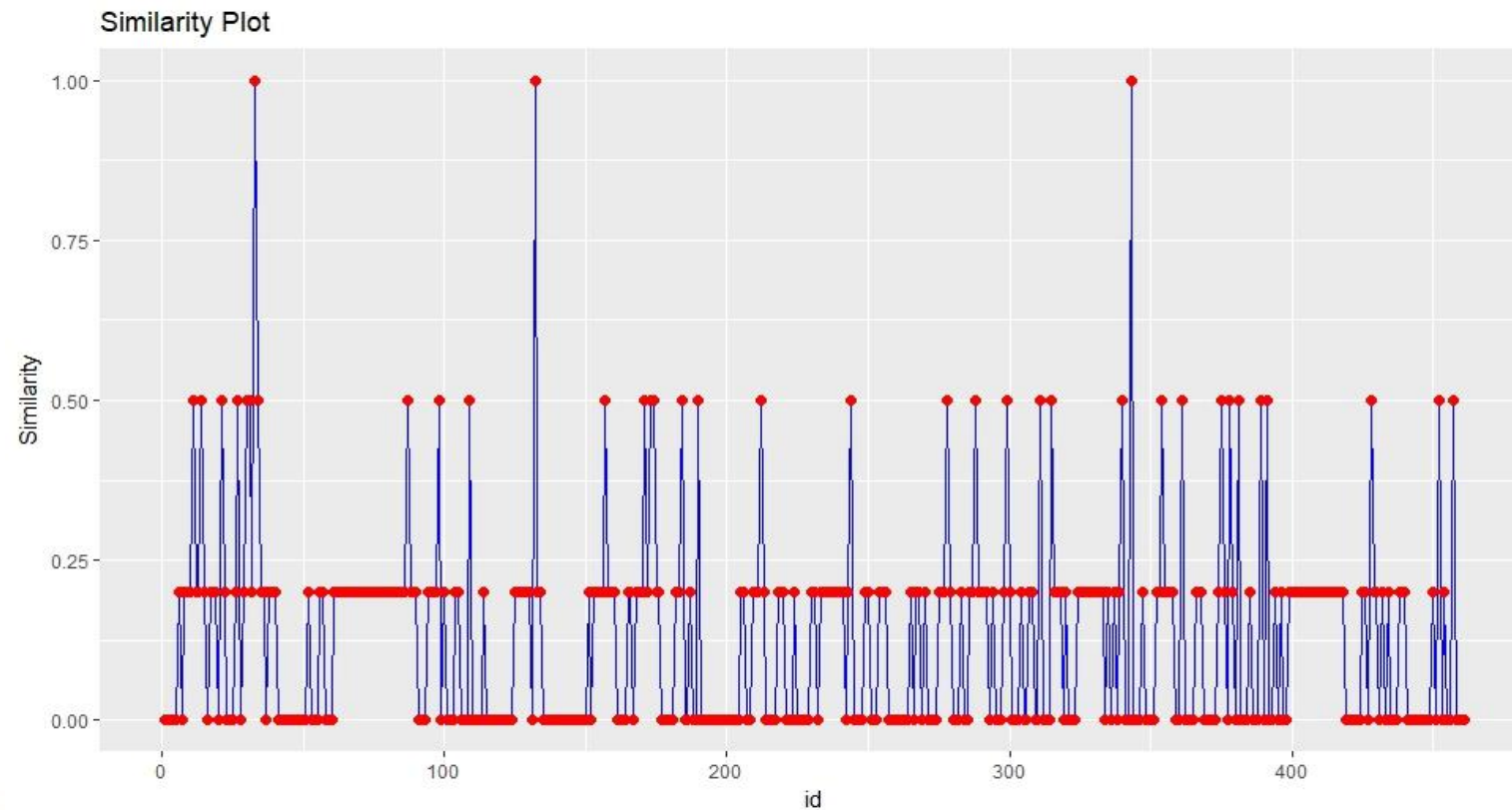
Example - 5 (id = 343, k = 50)

- KNN result

[illegible]

Example – 5 (id = 343, k = 50)

- KNN similarity plot



Example - 5 (id = 343, k = 50)

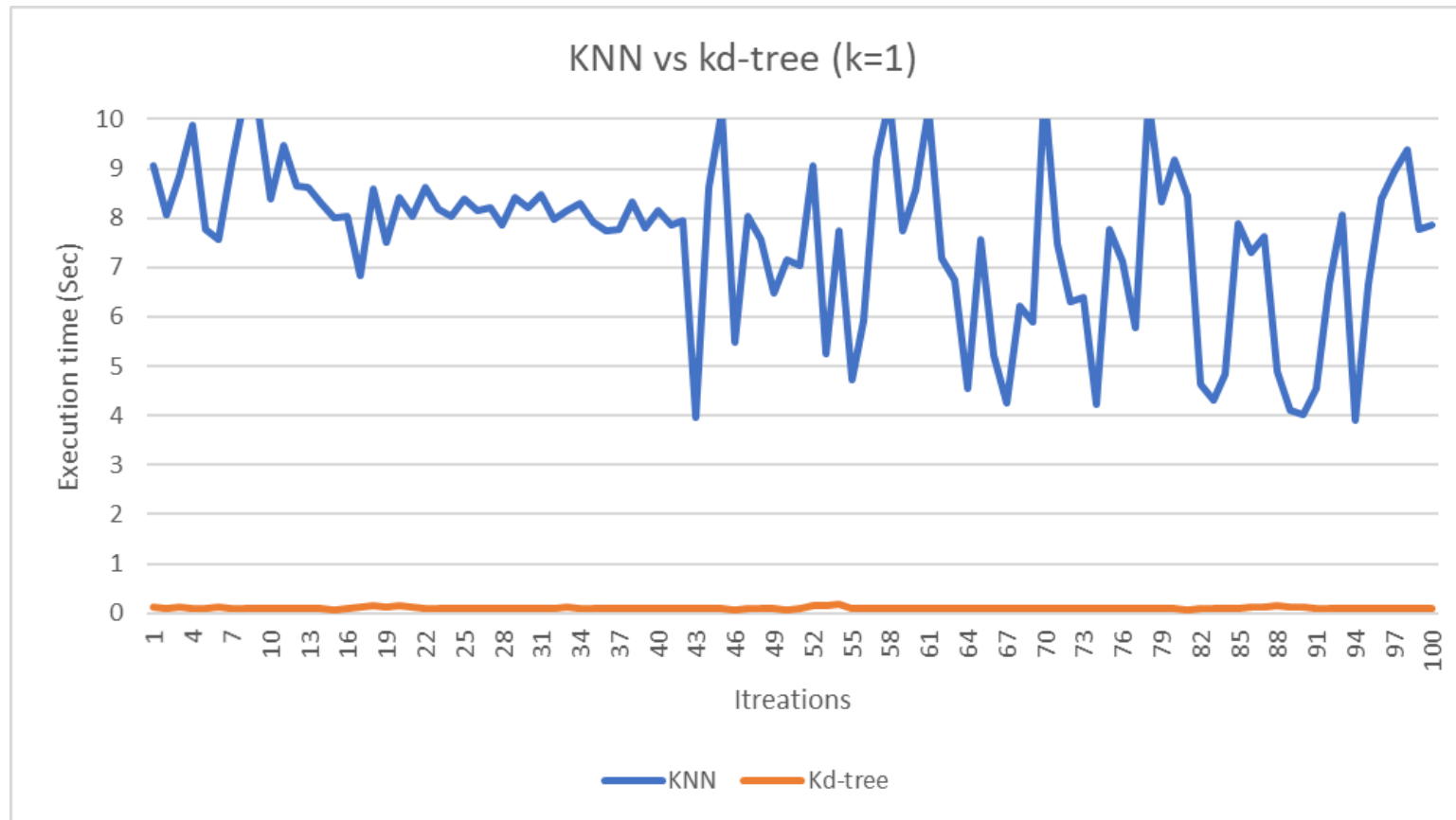
- Kd-tree result

[illegible]

KNN and kd-tree comparison

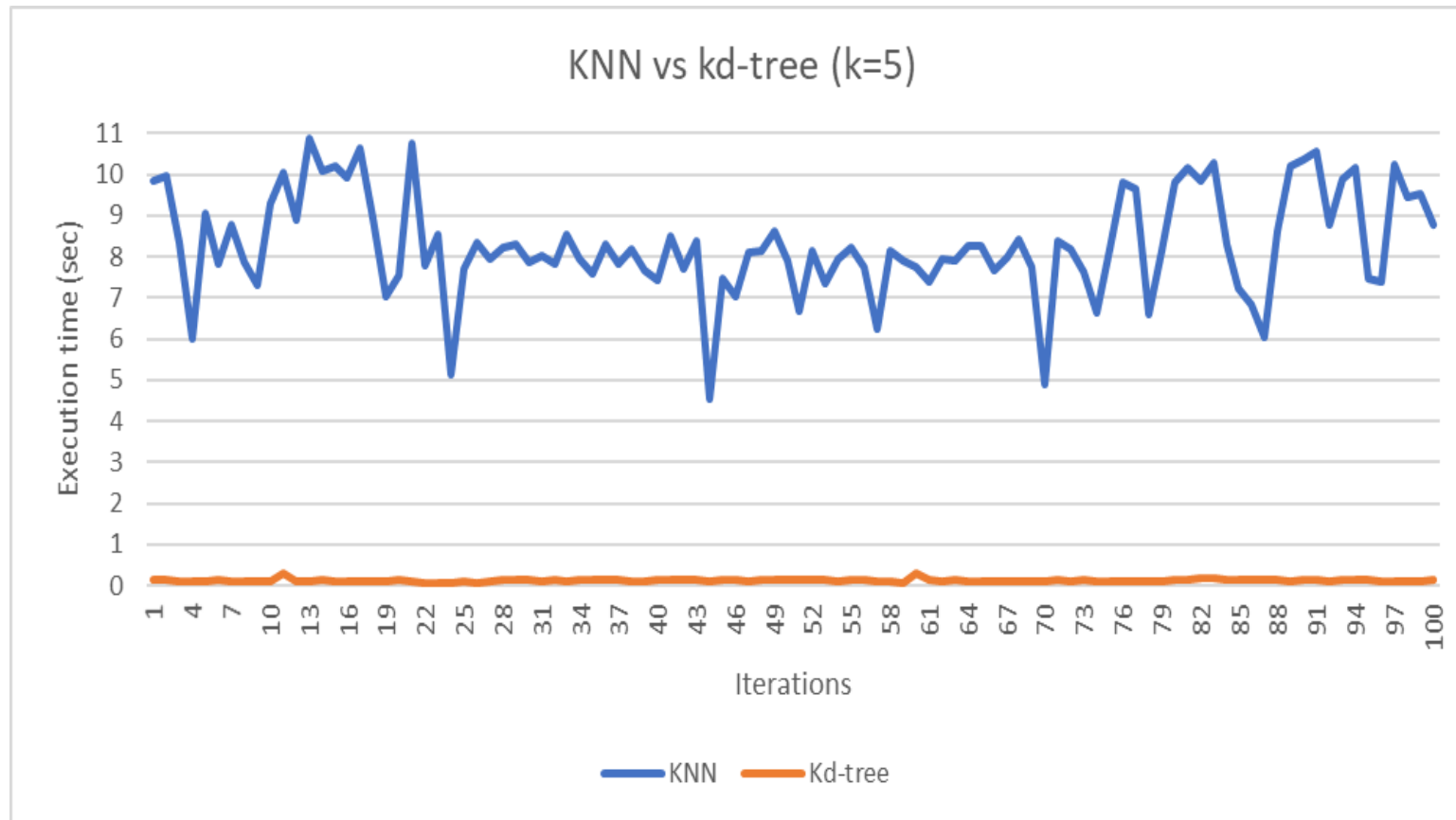
- KNN and kd-tree are compared based on execution time.
- KNN and kd-tree are compared for $k = 1, 5, 10, 20$ and 50 .
- 100 queries were executed for every value of k and average execution time was calculated.

KNN and kd-tree comparison (k = 1)



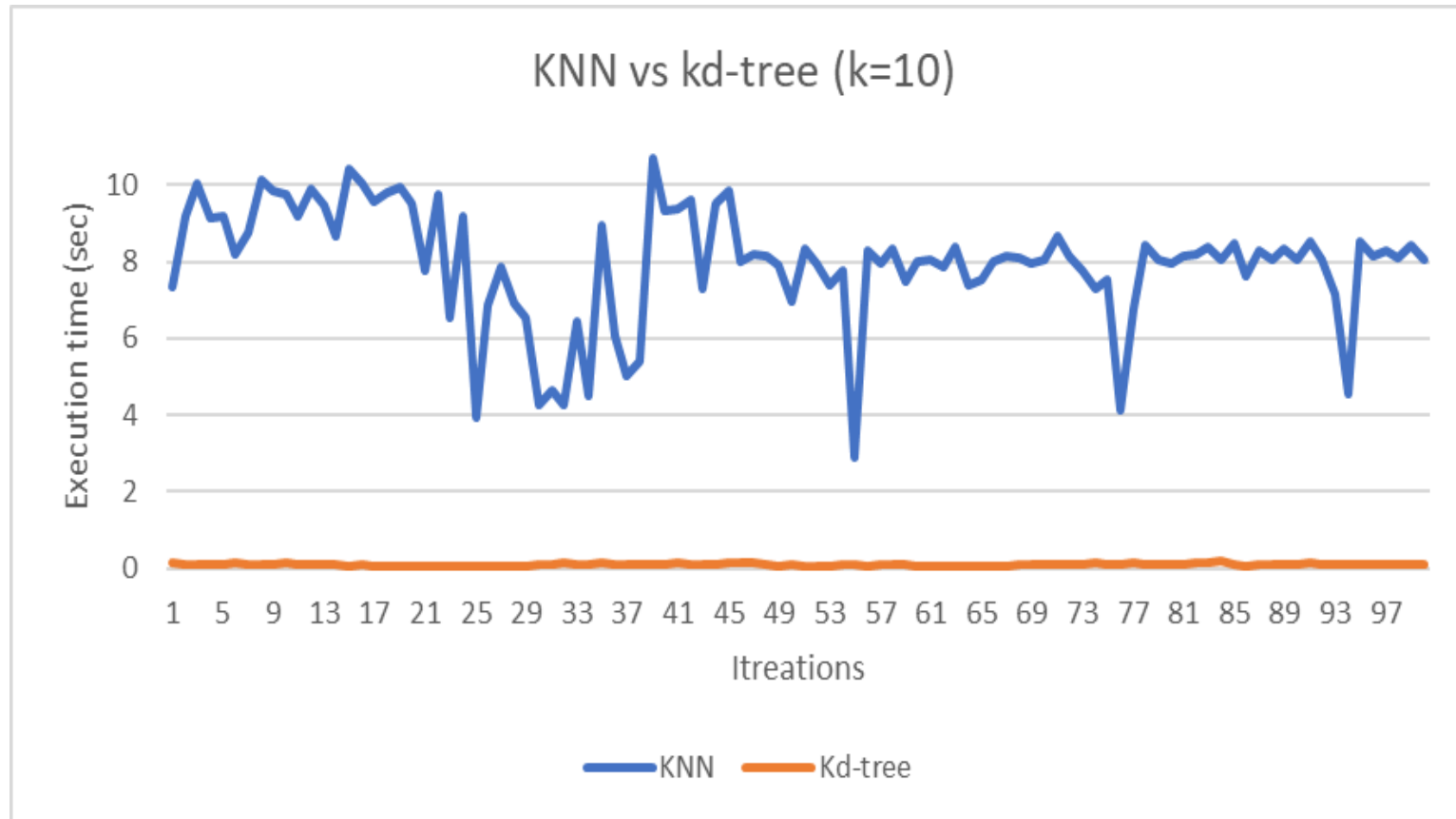
- Average execution time
 - KNN = 7.55 sec
 - Kd-tree = 0.10 sec

KNN and kd-tree comparison (k = 5)



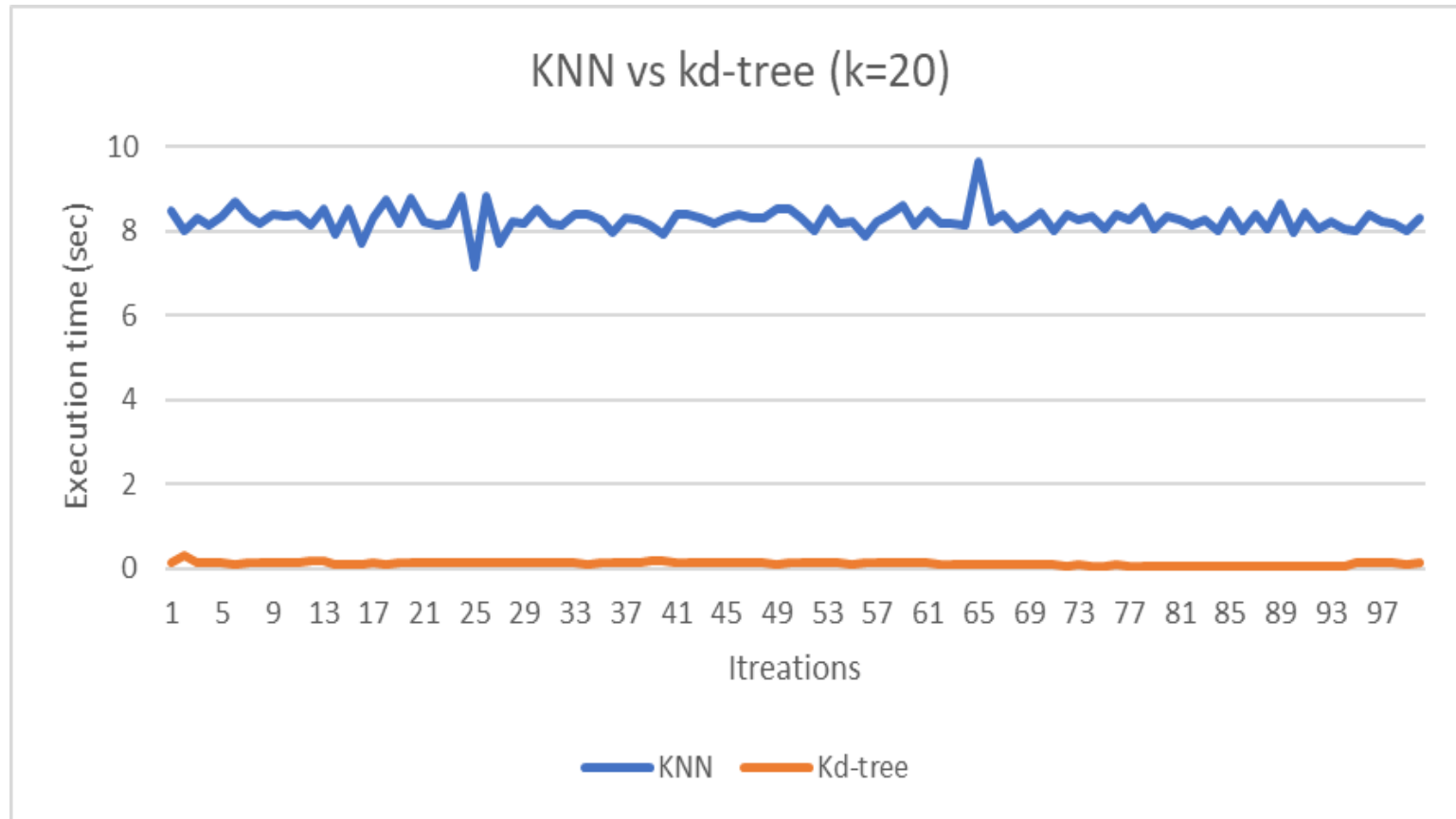
- Average execution time
 - KNN = 8.30 sec
 - Kd-tree = 0.12 sec

KNN and kd-tree comparison (k = 10)



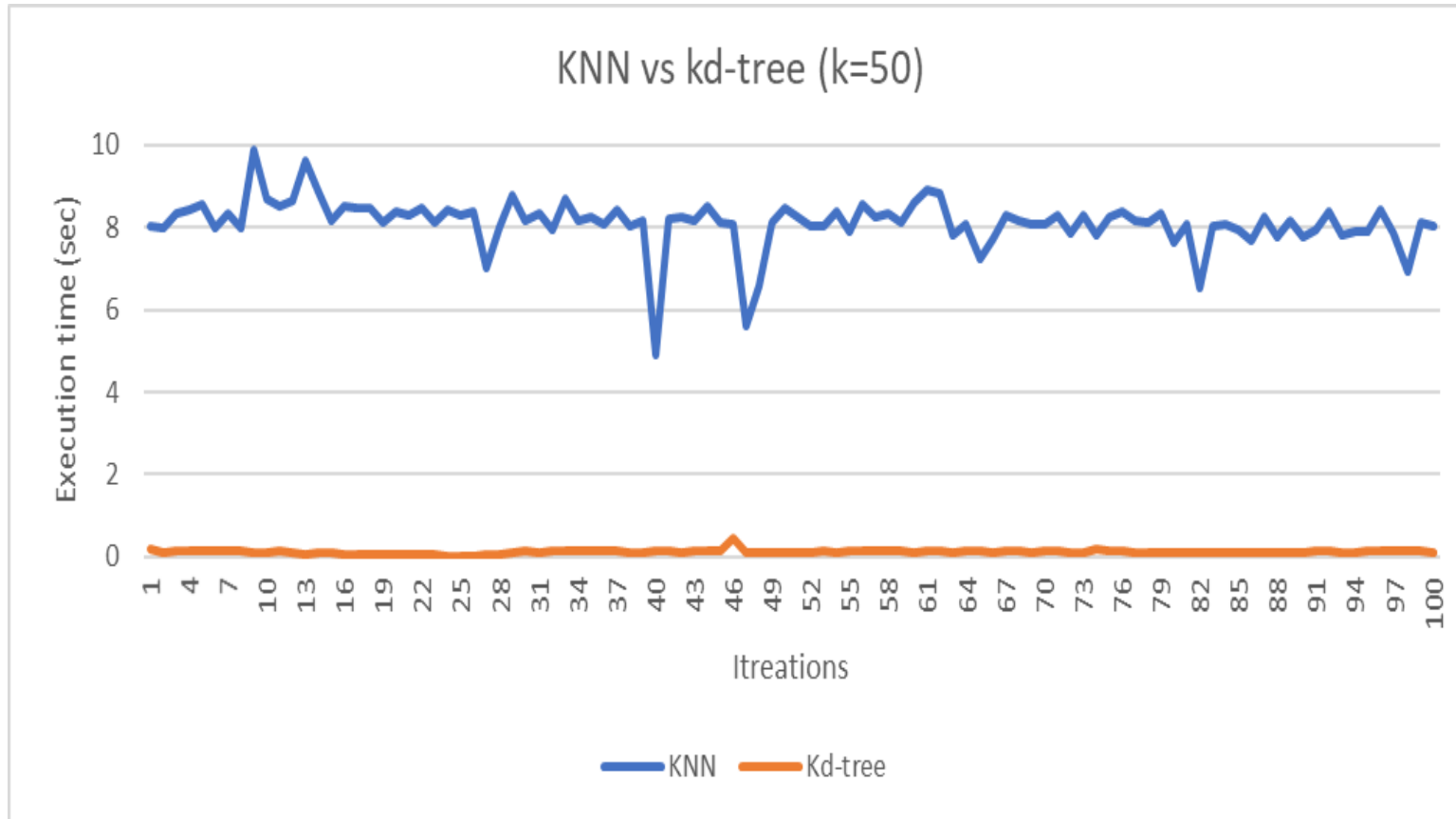
- Average execution time
 - KNN = 7.96 sec
 - Kd-tree = 0.11 sec

KNN and kd-tree comparison (k = 20)



- Average execution time
 - KNN = 8.27 sec
 - Kd-tree = 0.11 sec

KNN and kd-tree comparison (k = 50)



- Average execution time
 - KNN = 8.12 sec
 - Kd-tree = 0.12 sec

Conclusion

- Kd-tree is more accurate as compared to kNN.
- Kd-tree is efficient for execution as compared to kNN.
 - For $k = 1$ kd-tree is 74.5% faster than kNN
 - For $k = 5$ kd-tree is 68.16% faster than kNN
 - For $k = 10$ kd-tree is 71.36% faster than kNN
 - For $k = 20$ kd-tree is 74.18% faster than kNN
 - For $k = 50$ kd-tree is 66.66% faster than kNN

References

- <https://cran.r-project.org/web/packages/RMySQL/RMySQL.pdf>
- <https://cran.r-project.org/web/packages/proxy/proxy.pdf>
- <https://cran.r-project.org/web/packages/tictoc/tictoc.pdf>
- <https://cran.r-project.org/web/packages/rflann/rflann.pdf>
- <https://www.r-bloggers.com/accessing-mysql-through-r/>
- <https://stackoverflow.com/questions/24142576/one-hot-encoding-in-r-categorical-to-dummy-variables>
- <https://stackoverflow.com/questions/6262203/measuring-function-execution-time-in-r>



Thank You

