Experimental study of Descriptive and Inferential Analytics Approaches for Real Estate Buyers using ‘R’ Tools

Sameer Jain
Assistant Professor
National Institute of Construction Management and Research,
Pune, India

Abstract - In the building industry, analytics has altered the fundamental pattern of data processing and forecasting. Delights, the lobbying firm of a newly established real estate buyer, is aiming to penetrate the Melbourne property market. In order to generate insights on various aspects of this booming market, senior management is keen to capitalize on large volumes of historical real estate data. Insights can vary through location, seasonality, price patterns, area features, land features, property features and numerous other aspects. A massive dataset of real estate transactions in Melbourne, near 20,000 records from 2017, has been obtained by the company. Huge quantities of structured and unstructured data are generated in the industry and we will help a company make a game-changing decision with this knowledge. In this article, the descriptive and inferential analytics approach is used to extract insights into the real estate industry.

Keywords: Analytics; R; Real Estate; Inferential; Multiple Linear Regressions;

1.0 INTRODUCTION

As a scientist, you have to collect experiment data and analyze those as a part of the scientific method. In all those experiments, the data are very complicated and the understanding of that information is done by the help of graphs and statistical models. Graphics and modeling are done with the help of computer software and computing is just one of the skills necessary for a scientist. Training in the R needs lot of skills’’ is software used for analytics and it has the potential for statistical analysis of data. R has many important features like built-in functions libraries which have basic to advanced statistical functions. The R libraries installed from the CRAN mirror sites where R is available.

R has many data mining algorithms, as well as data visualization and data manipulation mechanisms. It covers all statistical applications from easy to complex and it would allow you to complete all your statistical training using R. Also, it covers topics in a consistent way so that the programming you learn for say linear models will also be done the same way for non-linear models, hence it will reduce your efforts. This consistency is convenient and also gives an understanding of statistical modeling. Modern statistics has simplified many problems through the use of graphics and computer intensive functions and tools.

2.0 LITERATURE REVIEW

In this to review the relevant literature, using analytics and big data for decision making on the basis of complex and large data sets which encompasses the tools for collecting and analyzing the structured and semi structured and unstructured data. R is a programming language and software environment for statistical computing and graphics supported by R foundation for statistical computing. The ‘R’ is widely used among statisticians and data miners for developing statistical software and data analysis [Wikipedia].

R is an implementation of S programming language created by two University of Auckland statisticians Ross Ihaka and Robert Gentlemen.

There are two analogies for the name ‘R’: (i) based on the first names of the two R author’s and (ii) As a play on the names of S (S programming language).

According to KDnuggests [11] online newsletter conducted polls in 2012, 2013 and 2014 asking the question “What statistics/programming languages you used for an analytics/ data mining/ data science work”. The result show that SQL, SAS, PYTHON, and R- hold a commanding lead and 91% of all respondents used one of them.

Big data analytics is a term through which real time data is analyze and managing both structured, semi-structured and unstructured data for decision making and optimize processes [7,8].

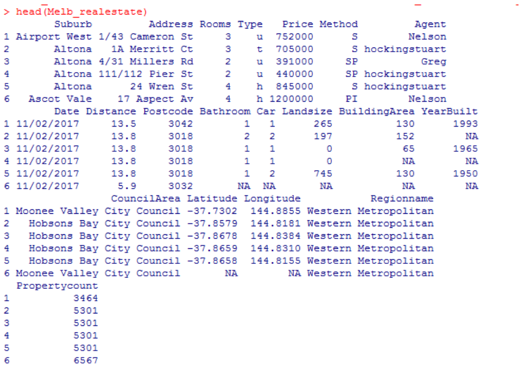
3.0 RESEARCH METHODOLOGY

Using descriptive research, statistical methodologies to analyze the factors that affect the real estate market in Melbourne are used by experiment-based 'R' tool to forecast the value of homes in Melbourne. This paper helps to provide insights into the real estate market, Domain sights, real estate firm needed move in Melbourne.

Step 1: Obtain the data
Set the working directory using setwd () command in the 'R' tool to your respective folder.

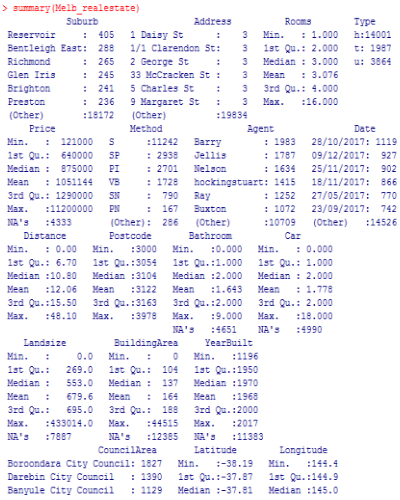
Load the dataset.

Melbourne _realestate = read.csv ("Melbournerealestate.csv")



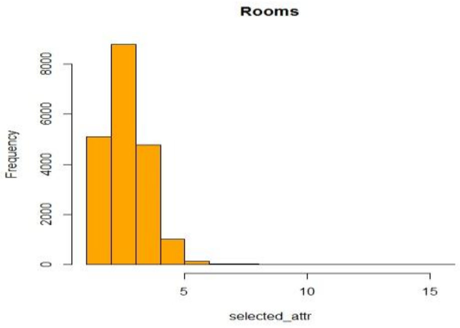
```
> head(Melb_realestate)
        Suburb          Address Rooms Type   Price Method         Agent
1 Airport West 1/43 Cameron St     3    u  752000      S        Nelson
2       Altona   1A Merritt Ct     3    t  705000      S hockingstuart
3       Altona 4/31 Millers Rd     2    u  391000     SP          Greg
4       Altona 111/112 Pier St     2    u  440000     SP hockingstuart
5       Altona      24 Wren St     4    h  845000      S hockingstuart
6   Ascot Vale    17 Aspect Av     4    h 1200000     PI        Nelson
        Date Distance Postcode Bathroom Car Landsize BuildingArea YearBuilt
1 11/02/2017     13.5     3042        1   1      265          130      1993
2 11/02/2017     13.8     3018        2   2      197          152        NA
3 11/02/2017     13.8     3018        1   1        0           65      1965
4 11/02/2017     13.8     3018        1   1        0           NA        NA
5 11/02/2017     13.8     3018        1   2      745          130      1950
6 11/02/2017      5.9     3032       NA  NA       NA           NA        NA
                 CouncilArea Latitude Longitude            Regionname
1 Moonee Valley City Council -37.7302  144.8855 Western Metropolitan
2   Hobsons Bay City Council -37.8579  144.8181 Western Metropolitan
3   Hobsons Bay City Council -37.8678  144.8384 Western Metropolitan
4   Hobsons Bay City Council -37.8659  144.8310 Western Metropolitan
5   Hobsons Bay City Council -37.8658  144.8155 Western Metropolitan
6 Moonee Valley City Council       NA        NA Western Metropolitan
  Propertycount
1          3464
2          5301
3          5301
4          5301
5          5301
6          6567
```

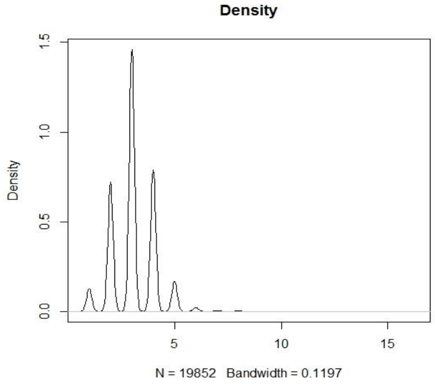
Step 2: Dataset Summary



```
> summary(Melb_realestate)
          Suburb                  Address           Rooms          Type
 Reservoir     :  405   1 Daisy St       :     3   Min.   : 1.000   h:14001
 Bentleigh East:  288   1/1 Clarendon St:     3   1st Qu.: 2.000   t: 1987
 Richmond      :  265   2 George St      :     3   Median : 3.000   u: 3864
 Glen Iris     :  245   33 McCracken St :     3   Mean   : 3.076
 Brighton      :  241   5 Charles St     :     3   3rd Qu.: 4.000
 Preston       :  236   9 Margaret St    :     3   Max.   :16.000
 (Other)       :18172   (Other)          :19834
     Price              Method              Agent                Date
 Min.   :  121000   S      :11242   Barry        : 1983   28/10/2017: 1119
 1st Qu.:  640000   SP     : 2938   Jellis       : 1787   09/12/2017:  927
 Median :  875000   PI     : 2701   Nelson       : 1634   25/11/2017:  902
 Mean   : 1051144   VB     : 1728   hockingstuart: 1415   18/11/2017:  866
 3rd Qu.: 1290000   SN     :  790   Ray          : 1252   27/05/2017:  770
 Max.   :11200000   PN     :  167   Buxton       : 1072   23/09/2017:  742
 NA's   :4333       (Other):  286   (Other)      :10709   (Other)   :14526
    Distance         Postcode        Bathroom          Car
 Min.   : 0.00   Min.   :3000   Min.   :0.000   Min.   : 0.000
 1st Qu.: 6.70   1st Qu.:3054   1st Qu.:1.000   1st Qu.: 1.000
 Median :10.80   Median :3104   Median :2.000   Median : 2.000
 Mean   :12.06   Mean   :3122   Mean   :1.643   Mean   : 1.778
 3rd Qu.:15.50   3rd Qu.:3163   3rd Qu.:2.000   3rd Qu.: 2.000
 Max.   :48.10   Max.   :3978   Max.   :9.000   Max.   :18.000
                                NA's   :4651    NA's   :4990
    Landsize          BuildingArea      YearBuilt
 Min.   :     0.0   Min.   :    0   Min.   :1196
 1st Qu.:   269.0   1st Qu.:  104   1st Qu.:1950
 Median :   553.0   Median :  137   Median :1970
 Mean   :   679.6   Mean   :  164   Mean   :1968
 3rd Qu.:   695.0   3rd Qu.:  188   3rd Qu.:2000
 Max.   :433014.0   Max.   :44515   Max.   :2017
 NA's   :7887       NA's   :12385   NA's   :11383
                    CouncilArea      Latitude        Longitude
 Boroondara City Council: 1827   Min.   :-38.19   Min.   :144.4
 Darebin City Council   : 1390   1st Qu.:-37.87   1st Qu.:144.9
 Banyule City Council   : 1129   Median :-37.81   Median :145.0
```

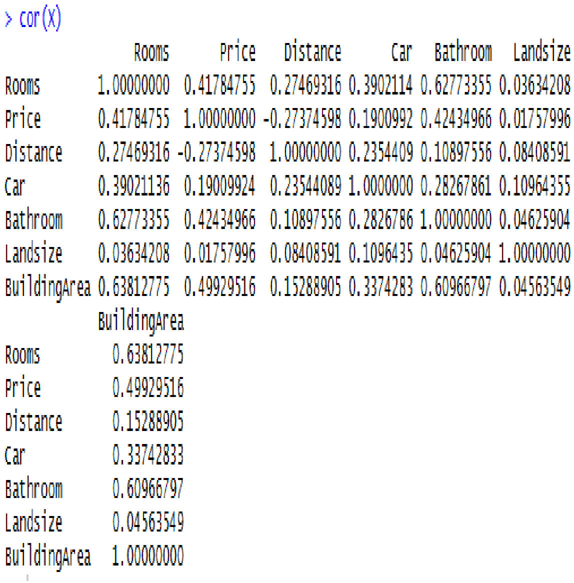
Step 3: Evaluate data distribution





Step 4: Reduce Data Dimensions

Analyze the correlation of attributes. Explore the correlation between attributes in the dataset.



```
> cor(X)
                 Rooms       Price    Distance       Car   Bathroom   Landsize
Rooms        1.00000000  0.41784755  0.27469316 0.3902114 0.62773355 0.03634208
Price        0.41784755  1.00000000 -0.27374598 0.1900992 0.42434966 0.01757996
Distance     0.27469316 -0.27374598  1.00000000 0.2354409 0.10897556 0.08408591
Car          0.39021136  0.19009924  0.23544089 1.0000000 0.28267861 0.10964355
Bathroom     0.62773355  0.42434966  0.10897556 0.2826786 1.00000000 0.04625904
Landsize     0.03634208  0.01757996  0.08408591 0.1096435 0.04625904 1.00000000
BuildingArea 0.63812775  0.49929516  0.15288905 0.3374283 0.60966797 0.04563549
             BuildingArea
Rooms          0.63812775
Price          0.49929516
Distance       0.15288905
Car            0.33742833
Bathroom       0.60966797
Landsize       0.04563549
BuildingArea   1.00000000
```

Plots to analyze the relationship:

As shown below, from the plot, the number of rooms is positively correlated with the price and the distance from the town of the property is negatively correlated.

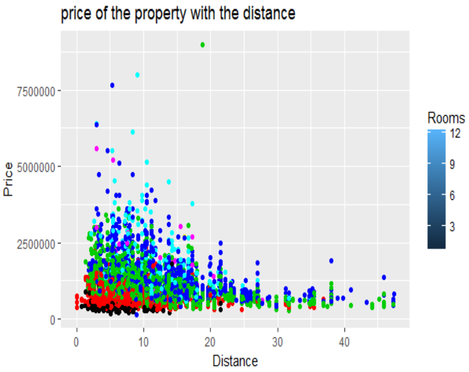
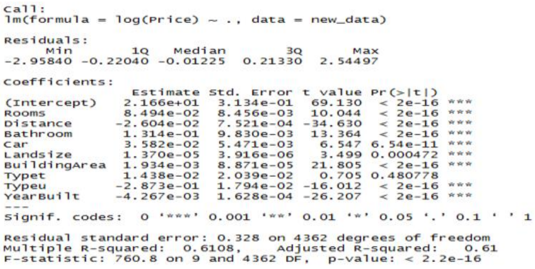


Figure 1: Distance and Price graph

<u>Multiple linear regression:</u>

To construct the relationship between the independent and dependent variables, multiple linear regression (MLR) models are used. The dependent variable is price, which can affect the remaining variables. Multiple linear regression helps to forecast continuous data, where the property's price can be predicted using the fitted multiple regression model.

Data Transformation:
We first read the data from the Melbourne real estate file data sets and generated a subset of numerical values to check for correlation. After that, the subset values are converted into the variables of the factor and then the multiple linear regression model is constructed. As the distribution of this variable equips the other data, the price variable is converted to logarithm.



```
Call:
lm(formula = log(Price) ~ ., data = new_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.95840 -0.22040 -0.01225  0.21330  2.54497 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.166e+01  3.134e-01  69.130  < 2e-16 ***
Rooms        8.494e-02  8.456e-03  10.044  < 2e-16 ***
Distance    -2.604e-02  7.521e-04 -34.630  < 2e-16 ***
Bathroom     1.314e-01  9.830e-03  13.364  < 2e-16 ***
Car          3.582e-02  5.471e-03   6.547 6.54e-11 ***
Landsize     1.370e-05  3.916e-06   3.499 0.000472 ***
BuildingArea 1.934e-03  8.871e-05  21.805  < 2e-16 ***
Typet        1.438e-02  2.039e-02   0.705 0.480778    
Typeu       -2.873e-01  1.794e-02 -16.012  < 2e-16 ***
YearBuilt   -4.267e-03  1.628e-04 -26.207  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.328 on 4362 degrees of freedom
Multiple R-squared:  0.6108,	Adjusted R-squared:   0.61 
F-statistic: 760.8 on 9 and 4362 DF,  p-value: < 2.2e-16
```

Interpretation of output

Significance of variables for predictor:

As p>.05, the Typet variable is insignificant; the p value greater than .05 is insignificant; thus, all the other variables except the Typet variable are important.

The adjusted R squared value reveals that this regression model explains 61 percent variability, rest due to randomness. The overall model is therefore critical since p<.05, with F-statistics 760.88.

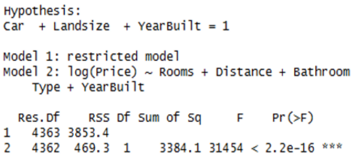
<u>Hypothesis testing:</u>

<u>*Scenario I:*</u>

<u>To verify that the property is more distance-biased than the number of rooms</u>

H0: To measure whether the price is equivalent to the distance of the property from the city.

H1: Alternate is the price is not equivalent to the distance of the property from the city.



```
Hypothesis:
Car  + Landsize  + YearBuilt = 1

Model 1: restricted model
Model 2: log(Price) ~ Rooms + Distance + Bathroom + Car + Landsize + BuildingArea + 
    Type + YearBuilt

  Res.Df    RSS Df Sum of Sq     F    Pr(>F)    
1   4363 3853.4                                 
2   4362  469.3  1    3384.1 31454 < 2.2e-16 ***
```

Interpretation:

The p value is less than .05, so there is no chance to accept the null hypothesis, it suggests

that the price is not identical and it differs, implying that a property that is 3BHK is overpriced

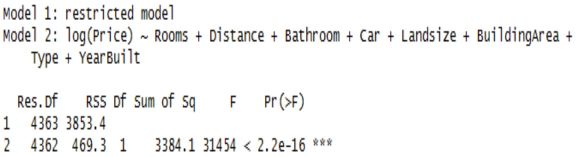
than a property near the city, located about 15-20 kms from the city.

<u>*Scenario 2:*</u>

<u>To check if the cost of the property is affected by the space of the car,
the year constructed, the size of the land and other variables.</u>

H0: To measure whether the price of the property is more influential than other
variables on car space, year built, land size.

H1: With these variables, the property price is not influential.



```
Model 1: restricted model
Model 2: log(Price) ~ Rooms + Distance + Bathroom + Car + Landsize + BuildingArea + 
    Type + YearBuilt

  Res.Df    RSS Df Sum of Sq     F    Pr(>F)    
1   4363 3853.4                                 
2   4362  469.3  1    3384.1 31454 < 2.2e-16 ***
```

The p value is lower than .05 from above, so the price of the property is more

affected by other variables.

<u>*Scenario 3:*</u>

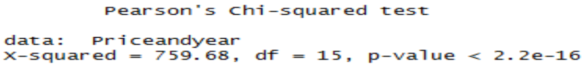
<u>To measure whether the year of completion of the property Affects the price of the roperty.</u>

H0: The property's price and year of building are independent.
H1: The year of construction determines the price of the property.

Data collection:

In this, we apply the chi-squared test of Pearson to verify the goodness of the
fit price value. Property prices are divided into five groups, ranging from cheap
to extremely expensive. Based on the year of completion, the year constructed is
clustered from the old building to the new.

Output:



```
        Pearson's Chi-squared test

data:  Priceandyear
X-squared = 759.68, df = 15, p-value < 2.2e-16
```

Interpretation:

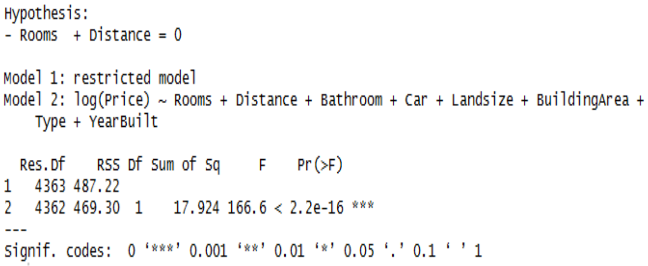
Therefore, from the above output p<0.05, it indicates that the price of the property is affected by the year constructed.

Scenario 4:

To determine if the number of spaces for cars depends on the distance from the city.

H0: A space for a car is independent of distance from the city.

H1: Car spaces are affected by distance.



```
Hypothesis:
- Rooms  + Distance = 0

Model 1: restricted model
Model 2: log(Price) ~ Rooms + Distance + Bathroom + Car + Landsize + BuildingArea +
    Type + YearBuilt

  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1   4363 487.22
2   4362 469.30  1    17.924 166.6 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpretation:

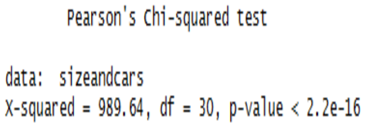
So, the number of car spaces depends on the distance from the city, from the above output p<0.05. The underlying hypothesis is that suburbs can have more car space than the city.

Scenario 5:

To determine if the number of car spaces is specified by the building size.

H0: Building size and space for cars are independent from each other.

H1: The amount of car space is influenced by the building's size.



```
        Pearson's Chi-squared test

data:  sizeandcars
X-squared = 989.64, df = 30, p-value < 2.2e-16
```

P<0.05, which means null hypothesis rejection, because the number of spaces in the car depends on the size of the property.

CONCLUSION

In this study, R was used because it is an open-source software that can also be used for all other paid packages, such as SPSS, MS EXCEL, Minitab, Tata, etc. Inferential, statistical and multiple linear regression techniques are extracted from the above research for the real estate sector. It shows that the price of Melbourne City property is affected by rooms, land size, year built, and distance from the city. The other variables, such as car space, type of house, have weak correlations and the space of the car depends on the size of the buildings. In this, there is a lot of space for predicting the dimensions and measuring by using data visualizations and classification techniques.

REFERENCES

[1] Abbott, D. (2014). Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst, Wiley.
[2] Boehmke, Bradley C. and Jackson, Ross A. (2016) “Unpacking the true cost of ‘free’ statistical software.” OR/MS Today, vol. 43, 26–27.
[3] Bradley Boehmke(2016): Data Wrangling with R, Springer.
[4] Deepali Arora, Piyush Malikanalytics (2015). Key To Go From Hoarding Big Data To Deriving Value, IEEE.
[5] Dhanya Jothimani, Ravi Shankar and Surendra S. Yadav (2014). A Big Data Analytical Framework for Portfolio Optimization.
[6] H. J Watson (2013), ‘’ All about Analytics ‘’, International Journal of Business Intelligence Research, Vol.4. No.2, pp. 13 – 28.
[7] H. Chen, R. H. L. Chiang, and V. C. Storey (2012), “Business Intelligence and Analytics: From Big Data to Big Impact,” MIS Q., vol. 36, no. 4, pp. 1165–1188.
[8] Ihaka, Ross, and Robert Gentleman. (1996) “R: A language for data analysis and graphics.” Journal of Computational and Graphical Statistics 5, 299–314.
[9] J.L.Bale, D.Bele, R.Pirnat, V.A.Loncaric(2015), ‘Business Intelligence In e-Learning’, Florence, Italy , pp 369- 375
[10] Krause, A. (2016). Reproducible research in real estate: A review and an example. Journal of Real Estate Practice and Education, 19(1), 69–85.

[11] Krause, A. & Lipscomb, C. A. (2016). The data preparation process in real estate: Guidance and review. Journal of Real Estate Practice and Education, 19(1), 15–42.

[12] Morandat, Floréal, Brandon Hill, Leo Osvald, and Jan Vitek. “Evaluating the design of the R language.”(2012) In European Conference on Object-Oriented Programming, pp. 104–131. Springer Berlin Heidelberg.

[13] Nelson, M. L. (2009). Data-driven science: A new paradigm? Educause Review, 44(4), 6–12.

[14] P. Russom“Big Data Analytics (2011), TDWI Best practices Report. Seattle: The Data warehousing,http://tdwi.org/research/2011/09/bestpractices-report-4-big-data-analytices.aspx.

[15] Pollack, R. D., Klimberg, R. K., and Boklage, S.H. (2015) “The true cost of ‘free’ statistical software.” OR/MS Today, vol. 42, 34–35.

[16] The R Project (2018) [Internet], R Project Avalible from http://www.r-project.org [Accessed April 2019]

[17] S. Miller, S. Lucas, L. Irakliotis, M. Ruppa, T.Carlson and B.Perlowitz. (2012), “Demystifying Big Data: A practical Guide to Transforming the Business of Government”, Washington: Tech America Foundation.

[18] Taylor, J, Decision management systems: A practical guide to using business rules and predictive analytics. New York, NY: IBM Press.

[19] Yongho Ko And Seungwoo Han (2015). Big Data Analysis Based Practical Applications in Construction, Int'l Conf. on Advances in Big Data Analytics, pp.121-122.

[20] Internet Source / Web Source

[21] [a] www.r-project.org, accessed on 30 November 2020.

[22] [b] www.python.org/doc/essays/blurb/, accessed on 16 November 2020.

[23] [c] www.scala-lang.org/, accessed on 15 October 2020.

[24] [d] https://spark.apache.org/, accessed on 22 October 2020.

[25] [e] https://cran.r-project.org/web/packages/sparklyr/index.html, accessed on 25 December