Exam Template: Statistical Inference

21045690

Jan 2022: Sep21 run

# Instructions to students

You should only use the file Exam\_template.Rmd provided on blackboard and you should load this file from your scripts folder / directory.

Save this template as your studentID.Rmd; you will upload this file as your submission. Change the information on line 3 of this file – changing the author information to your **student ID**. Do not change the authorship to your name.

Ensure that you save your data into your data folder (as discussed in class). You may use the files mypackages.R and helperFunctions.R from blackboard. If you use these files, do not alter them. If you wish to create additional files for custom functions that you have prepared in advance, make sure that you upload these in addition to your .Rmd file and your compiled output file.

Your should knit this file to a document **Word** format.

Any changes that you make to the data (e.g. variable name changes) should be made entirely within R.

The subsubsections labelled **Answer:** indicate where you should put in your written Answers. The template also provides blank code chunks for you to complete your Answers; you may choose to add additional chunks if required.

# load dataset   
ola\_cars <- read.csv("Jan\_2022\_Exam\_Data.csv")  
str(ola\_cars)

'data.frame': 52533 obs. of 10 variables:  
 $ brand : chr "Audi" "Audi" "Audi" "Audi" ...  
 $ model : chr " A1" " A6" " A1" " A4" ...  
 $ year : int 2017 2016 2016 2017 2019 2016 2016 2016 2015 2016 ...  
 $ price : int 12500 16500 11000 16800 17300 13900 13250 11750 10200 12000 ...  
 $ transmission: chr "Manual" "Automatic" "Manual" "Automatic" ...  
 $ mileage : int 15735 36203 29946 25952 1998 32260 76788 75185 46112 22451 ...  
 $ fuelType : chr "Petrol" "Diesel" "Petrol" "Diesel" ...  
 $ tax : int 150 20 30 145 145 30 30 20 20 30 ...  
 $ mpg : num 55.4 64.2 55.4 67.3 49.6 58.9 61.4 70.6 60.1 55.4 ...  
 $ engineSize : num 1.4 2 1.4 2 1 1.4 2 2 1.4 1.4 ...

# Data description

This dataset is part of a larger dataset that has been collected to help to estimate the price of used cars.

It contains the following variables:

* brand (manufacturer)
* model (of car)
* year (of registration of the car)
* price (in GB pounds)
* transmission (type of gearbox)
* mileage (total distance covered by the car)
* fuelType (type of fuel used by the car)
* tax (annual cost of vehicle tax)
* mpg (miles per gallon - a measure of fuel efficiency)
* engineSize (size of the engine in litres)

# Question 1: Data Preparation (11 marks)

You are interested in modelling the price of vehicles that have all of the following properties:

* mileage less than 60000
* Manual transmission
* Petrol engine (fuelType)
* Costing less than £200 in annual Vehicle Tax.

Once you have selected the rows of data with these properties, then you must *use your studentID* to select a random sample of 2000 rows of the data to perform the rest of your analysis with.

You should remove any redundant variables (where only one value remains in that variable).

This subset of the data is what you should use for the rest of this assessment.

1. Explain what data preparation is required in order for the data in Jan\_2022\_Exam\_Data.csv to be suitable for this analysis.

**(4 marks)**

### Answer:

Firstly, we need to select the rows that satisfies the conditions we are interested in modelling which are mileage less than 60000, manual transmission, petrol engine (fuel Type), costing less than £200 in annual vehicle tax and save it as a new data frame.

Secondly, set seed to my student id 21045690 to get 2000 data as a sample size from the selected rows satisfies the conditions met. The next step is to clean the data i.e remove any variables that have no effect on the the selected data frame.

1. Implement the required data preparation in the code chunk below:

**(7 marks)**

### Answer:

# select the rows that satisfies the condition  
ola\_cars\_mod <- subset(ola\_cars, ola\_cars$mileage < 60000 & ola\_cars$transmission == "Manual" & ola\_cars$fuelType == "Petrol" & ola\_cars$tax < 200,)  
  
# set seed to my student id and select 2000 rows from the subset data  
set.seed(21045690)  
no\_of\_rows <- nrow(ola\_cars\_mod)   
  
# ind is the observation rows sampled  
ind <- sample(1:no\_of\_rows, size = 2000, replace = FALSE)   
  
# sort the observation of row sampled  
obs\_ola\_cars <- sort(ind)  
obs\_ola\_cars

[1] 1 2 16 25 29 34 49 52 57 62 71 72  
 [13] 80 94 97 98 108 118 156 158 161 183 185 191  
 [25] 192 195 196 199 202 204 206 221 228 230 232 233  
 [37] 240 243 258 274 282 283 284 291 309 316 324 325  
 [49] 335 337 355 368 370 371 373 375 376 378 381 386  
 [61] 387 391 393 395 396 409 419 423 429 433 441 445  
 [73] 448 454 465 466 468 473 476 481 492 495 503 511  
 [85] 517 535 540 559 567 590 594 599 609 617 624 627  
 [97] 628 645 648 652 659 662 665 666 668 675 678 697  
 [109] 699 701 703 709 724 730 740 742 744 750 761 762  
 [121] 779 780 789 815 816 817 818 827 828 835 838 841  
 [133] 849 851 863 877 886 892 903 909 910 912 914 921  
 [145] 947 949 950 952 961 969 973 990 994 997 1003 1005  
 [157] 1017 1021 1029 1030 1045 1056 1061 1062 1070 1071 1076 1078  
 [169] 1089 1098 1110 1115 1123 1124 1143 1147 1157 1161 1169 1174  
 [181] 1191 1193 1194 1202 1213 1220 1222 1225 1227 1239 1242 1246  
 [193] 1252 1259 1262 1265 1266 1281 1284 1300 1311 1313 1314 1317  
 [205] 1330 1331 1340 1343 1350 1359 1364 1373 1389 1405 1409 1411  
 [217] 1418 1425 1434 1448 1450 1461 1470 1471 1474 1477 1480 1494  
 [229] 1500 1502 1503 1507 1512 1514 1521 1535 1539 1549 1565 1571  
 [241] 1572 1581 1585 1598 1604 1614 1617 1620 1625 1634 1643 1644  
 [253] 1651 1658 1666 1675 1685 1691 1693 1695 1701 1710 1718 1730  
 [265] 1749 1757 1765 1767 1775 1778 1781 1787 1801 1816 1817 1834  
 [277] 1881 1882 1887 1893 1901 1911 1916 1918 1940 1944 1948 1950  
 [289] 1951 1952 1955 1959 1966 1971 1974 1977 1979 1985 1988 1990  
 [301] 1999 2004 2005 2012 2020 2022 2030 2037 2047 2061 2063 2070  
 [313] 2077 2079 2110 2121 2129 2140 2142 2149 2161 2166 2181 2189  
 [325] 2193 2211 2214 2224 2227 2228 2232 2234 2236 2237 2238 2263  
 [337] 2274 2278 2281 2291 2294 2296 2305 2317 2325 2333 2334 2342  
 [349] 2348 2351 2363 2364 2365 2366 2376 2378 2381 2387 2391 2392  
 [361] 2413 2428 2430 2435 2452 2455 2473 2477 2494 2498 2501 2508  
 [373] 2510 2522 2526 2528 2530 2545 2549 2552 2555 2565 2566 2589  
 [385] 2591 2599 2602 2614 2615 2625 2642 2648 2656 2659 2663 2667  
 [397] 2668 2675 2676 2691 2692 2693 2704 2705 2707 2709 2713 2721  
 [409] 2729 2730 2732 2739 2741 2745 2747 2757 2769 2773 2775 2780  
 [421] 2789 2807 2809 2818 2834 2835 2836 2852 2858 2862 2865 2868  
 [433] 2870 2872 2900 2902 2922 2934 2936 2937 2939 2944 2949 2962  
 [445] 2965 2970 2976 2979 2981 3015 3020 3026 3032 3035 3042 3055  
 [457] 3060 3070 3095 3100 3102 3109 3110 3116 3141 3142 3147 3155  
 [469] 3174 3201 3204 3212 3218 3226 3227 3235 3244 3246 3249 3252  
 [481] 3256 3258 3270 3282 3286 3297 3306 3307 3313 3337 3339 3345  
 [493] 3352 3357 3360 3363 3366 3367 3384 3389 3407 3412 3438 3443  
 [505] 3445 3456 3470 3471 3475 3485 3495 3503 3506 3521 3528 3529  
 [517] 3534 3535 3536 3552 3559 3564 3570 3572 3580 3592 3606 3609  
 [529] 3625 3627 3628 3630 3634 3643 3665 3669 3675 3694 3695 3706  
 [541] 3708 3709 3716 3719 3729 3747 3758 3765 3771 3781 3788 3795  
 [553] 3801 3805 3809 3822 3827 3834 3842 3851 3857 3859 3863 3867  
 [565] 3877 3885 3887 3889 3892 3895 3901 3902 3910 3913 3931 3934  
 [577] 3936 3938 3947 3960 3961 3963 3967 3968 3974 3981 3994 3997  
 [589] 4001 4005 4012 4027 4028 4054 4055 4057 4064 4070 4073 4074  
 [601] 4084 4090 4101 4102 4103 4121 4125 4131 4133 4139 4145 4147  
 [613] 4157 4166 4172 4180 4197 4211 4213 4229 4248 4249 4259 4261  
 [625] 4265 4266 4268 4276 4281 4284 4300 4307 4315 4318 4327 4334  
 [637] 4336 4343 4348 4353 4357 4358 4362 4364 4373 4376 4382 4388  
 [649] 4389 4394 4395 4400 4405 4413 4431 4432 4444 4460 4468 4471  
 [661] 4476 4485 4486 4521 4526 4527 4530 4534 4535 4554 4571 4573  
 [673] 4578 4584 4591 4614 4619 4627 4637 4641 4643 4644 4650 4654  
 [685] 4657 4664 4667 4668 4671 4674 4682 4685 4697 4705 4709 4712  
 [697] 4722 4728 4731 4739 4743 4744 4750 4759 4760 4772 4778 4806  
 [709] 4809 4816 4819 4823 4831 4833 4841 4842 4852 4853 4855 4856  
 [721] 4859 4870 4886 4892 4897 4901 4903 4910 4911 4918 4921 4933  
 [733] 4944 4973 4976 4983 4988 4991 5001 5003 5017 5020 5021 5032  
 [745] 5034 5047 5056 5057 5062 5067 5070 5080 5094 5096 5098 5116  
 [757] 5117 5123 5130 5131 5148 5151 5152 5197 5201 5212 5216 5220  
 [769] 5241 5252 5253 5258 5259 5274 5276 5286 5290 5332 5334 5336  
 [781] 5338 5360 5363 5366 5367 5400 5418 5423 5426 5431 5435 5436  
 [793] 5442 5447 5458 5460 5464 5467 5473 5474 5481 5488 5498 5507  
 [805] 5515 5523 5524 5527 5535 5540 5551 5553 5556 5581 5586 5588  
 [817] 5602 5603 5608 5617 5620 5627 5650 5666 5669 5670 5685 5688  
 [829] 5690 5704 5709 5712 5728 5729 5736 5739 5741 5742 5743 5753  
 [841] 5783 5784 5785 5790 5807 5808 5812 5816 5820 5825 5834 5836  
 [853] 5837 5841 5843 5851 5852 5862 5863 5864 5868 5869 5873 5879  
 [865] 5883 5885 5939 5945 5967 5970 5974 5980 5987 5992 6001 6006  
 [877] 6009 6010 6012 6014 6019 6022 6029 6050 6051 6063 6074 6086  
 [889] 6094 6101 6106 6116 6119 6128 6134 6140 6150 6155 6157 6162  
 [901] 6167 6177 6185 6187 6196 6198 6202 6205 6211 6216 6231 6244  
 [913] 6245 6248 6250 6259 6266 6268 6285 6292 6295 6309 6321 6326  
 [925] 6340 6353 6360 6362 6364 6369 6379 6386 6394 6399 6405 6412  
 [937] 6416 6430 6432 6434 6435 6436 6450 6451 6457 6462 6474 6481  
 [949] 6484 6485 6487 6490 6491 6492 6502 6503 6506 6515 6517 6520  
 [961] 6534 6538 6543 6572 6578 6580 6587 6593 6600 6602 6605 6606  
 [973] 6619 6620 6628 6637 6643 6645 6647 6651 6662 6663 6675 6679  
 [985] 6683 6684 6689 6697 6699 6705 6710 6713 6726 6731 6737 6757  
 [997] 6758 6759 6761 6773 6776 6780 6801 6807 6809 6815 6835 6844  
[1009] 6845 6850 6856 6860 6871 6875 6882 6889 6890 6892 6905 6910  
[1021] 6920 6931 6944 6949 6959 6961 6974 6976 6981 6983 6990 7011  
[1033] 7013 7017 7019 7026 7029 7035 7049 7050 7063 7070 7071 7079  
[1045] 7093 7106 7107 7111 7116 7123 7124 7130 7139 7154 7159 7163  
[1057] 7178 7184 7189 7193 7215 7233 7236 7240 7242 7248 7251 7253  
[1069] 7256 7261 7268 7274 7278 7287 7303 7309 7311 7315 7318 7321  
[1081] 7325 7343 7347 7349 7360 7366 7369 7381 7384 7386 7395 7399  
[1093] 7402 7404 7406 7410 7411 7413 7417 7431 7437 7445 7476 7489  
[1105] 7493 7498 7520 7521 7522 7536 7537 7543 7557 7561 7566 7580  
[1117] 7581 7601 7605 7615 7616 7626 7647 7649 7655 7663 7679 7711  
[1129] 7725 7728 7730 7754 7756 7773 7784 7789 7800 7803 7824 7825  
[1141] 7830 7831 7836 7840 7842 7843 7855 7866 7896 7899 7909 7912  
[1153] 7927 7933 7936 7937 7939 7944 7945 7946 7951 7962 7970 7975  
[1165] 7979 7993 7995 7997 8000 8005 8006 8015 8018 8020 8029 8031  
[1177] 8042 8053 8055 8056 8074 8075 8086 8087 8096 8106 8115 8124  
[1189] 8127 8146 8149 8160 8165 8167 8171 8176 8179 8181 8182 8202  
[1201] 8204 8207 8209 8248 8253 8259 8268 8270 8272 8287 8298 8302  
[1213] 8313 8317 8323 8325 8337 8352 8365 8366 8368 8372 8373 8380  
[1225] 8382 8392 8397 8401 8404 8416 8418 8424 8425 8427 8433 8442  
[1237] 8446 8449 8454 8466 8471 8483 8488 8489 8490 8495 8502 8504  
[1249] 8507 8510 8524 8525 8540 8547 8565 8574 8587 8590 8591 8593  
[1261] 8594 8601 8605 8607 8611 8619 8628 8653 8658 8659 8665 8670  
[1273] 8675 8680 8682 8683 8686 8690 8695 8698 8705 8713 8714 8723  
[1285] 8731 8760 8776 8779 8785 8795 8797 8801 8802 8803 8812 8822  
[1297] 8849 8850 8851 8853 8858 8862 8881 8900 8924 8933 8938 8939  
[1309] 8944 8950 8952 8955 8957 8960 8961 8967 8981 8984 9002 9003  
[1321] 9019 9039 9042 9043 9045 9050 9059 9065 9070 9082 9096 9099  
[1333] 9115 9118 9121 9122 9126 9129 9130 9138 9139 9144 9145 9175  
[1345] 9180 9186 9201 9205 9211 9212 9222 9225 9231 9237 9242 9244  
[1357] 9246 9259 9267 9270 9272 9276 9283 9304 9312 9315 9316 9326  
[1369] 9329 9339 9351 9356 9373 9390 9391 9403 9407 9414 9417 9422  
[1381] 9426 9433 9434 9445 9446 9451 9453 9457 9458 9460 9500 9502  
[1393] 9509 9526 9528 9538 9542 9548 9551 9567 9570 9584 9590 9595  
[1405] 9611 9623 9631 9653 9664 9672 9676 9683 9689 9690 9692 9693  
[1417] 9699 9700 9709 9711 9719 9724 9728 9732 9736 9740 9743 9748  
[1429] 9749 9759 9767 9775 9780 9782 9790 9792 9806 9808 9810 9818  
[1441] 9820 9821 9827 9832 9834 9844 9845 9848 9855 9859 9867 9873  
[1453] 9882 9884 9888 9897 9899 9902 9907 9913 9920 9921 9922 9925  
[1465] 9927 9933 9935 9947 9956 9968 9973 9977 9982 9995 10003 10005  
[1477] 10017 10022 10029 10030 10048 10049 10070 10072 10086 10090 10104 10115  
[1489] 10120 10124 10129 10137 10140 10141 10146 10158 10160 10162 10174 10190  
[1501] 10204 10211 10218 10220 10231 10240 10251 10263 10267 10289 10292 10299  
[1513] 10300 10305 10313 10314 10330 10332 10346 10350 10359 10370 10372 10376  
[1525] 10382 10387 10393 10398 10407 10414 10419 10435 10441 10446 10448 10451  
[1537] 10462 10477 10482 10483 10486 10488 10490 10492 10508 10521 10525 10528  
[1549] 10530 10534 10550 10551 10557 10558 10564 10565 10566 10567 10573 10580  
[1561] 10583 10584 10586 10587 10598 10604 10607 10610 10612 10613 10617 10629  
[1573] 10631 10654 10665 10674 10679 10688 10692 10696 10698 10699 10702 10713  
[1585] 10714 10715 10716 10718 10719 10728 10730 10733 10735 10741 10745 10746  
[1597] 10749 10762 10763 10766 10777 10795 10796 10797 10800 10811 10813 10816  
[1609] 10831 10842 10856 10866 10873 10874 10885 10890 10895 10896 10903 10911  
[1621] 10912 10913 10919 10929 10932 10938 10957 10958 10959 10963 10971 10972  
[1633] 10977 10981 10983 10986 10995 10996 10998 10999 11003 11004 11008 11023  
[1645] 11036 11051 11053 11054 11062 11071 11085 11105 11107 11108 11111 11120  
[1657] 11123 11125 11132 11141 11163 11173 11190 11194 11207 11209 11217 11243  
[1669] 11259 11264 11277 11284 11287 11292 11299 11308 11309 11334 11336 11344  
[1681] 11348 11351 11355 11358 11359 11366 11377 11413 11414 11423 11425 11427  
[1693] 11430 11438 11442 11446 11451 11461 11463 11474 11479 11482 11493 11497  
[1705] 11504 11512 11516 11517 11530 11532 11533 11534 11541 11550 11556 11559  
[1717] 11561 11566 11567 11577 11582 11596 11603 11607 11630 11631 11633 11637  
[1729] 11638 11640 11647 11648 11650 11651 11664 11671 11683 11696 11706 11707  
[1741] 11728 11730 11732 11733 11740 11748 11751 11754 11766 11772 11785 11786  
[1753] 11792 11795 11798 11801 11809 11815 11821 11823 11831 11833 11850 11865  
[1765] 11875 11879 11886 11892 11894 11901 11914 11922 11936 11938 11941 11960  
[1777] 11965 11966 11971 11972 11973 11988 11998 12001 12002 12004 12015 12025  
[1789] 12026 12032 12035 12038 12042 12053 12056 12059 12068 12071 12073 12086  
[1801] 12099 12102 12106 12107 12119 12128 12129 12147 12159 12169 12176 12178  
[1813] 12179 12182 12186 12187 12193 12194 12195 12201 12216 12270 12274 12276  
[1825] 12286 12290 12299 12312 12314 12317 12324 12335 12337 12349 12364 12371  
[1837] 12377 12379 12382 12383 12388 12392 12398 12406 12411 12418 12426 12427  
[1849] 12428 12434 12446 12460 12462 12467 12468 12476 12481 12482 12488 12492  
[1861] 12493 12507 12511 12514 12515 12517 12521 12561 12571 12575 12577 12589  
[1873] 12632 12637 12641 12684 12697 12704 12706 12717 12729 12730 12735 12743  
[1885] 12759 12778 12780 12786 12787 12788 12796 12797 12811 12813 12815 12829  
[1897] 12834 12835 12841 12850 12851 12852 12856 12857 12894 12899 12902 12903  
[1909] 12905 12912 12915 12919 12923 12926 12931 12932 12939 12947 12952 12962  
[1921] 12972 12979 12981 13007 13021 13024 13034 13044 13045 13053 13077 13082  
[1933] 13090 13095 13098 13105 13125 13127 13131 13136 13139 13144 13146 13148  
[1945] 13155 13171 13180 13207 13213 13232 13237 13241 13245 13254 13258 13259  
[1957] 13261 13262 13268 13270 13274 13283 13284 13291 13305 13319 13321 13332  
[1969] 13335 13339 13343 13349 13351 13355 13362 13363 13371 13372 13382 13391  
[1981] 13422 13429 13458 13461 13471 13474 13499 13502 13504 13512 13533 13542  
[1993] 13545 13547 13550 13552 13556 13572 13575 13582

# filter ola\_cars\_mod data frame with the sorted observation row sampled to get 2000 rows of data  
cars\_by\_ola <- ola\_cars\_mod[obs\_ola\_cars,]   
view(cars\_by\_ola)  
  
# remove redundant variables by selecting variable  
view(cars\_by\_ola)   
cars\_df <- cars\_by\_ola[, c(1,2,3,4,6,8:10)]  
head(cars\_df,10)

brand model year price mileage tax mpg engineSize  
1 Audi A1 2017 12500 15735 150 55.4 1.4  
3 Audi A1 2016 11000 29946 30 55.4 1.4  
68 Audi A1 2017 11700 19268 150 67.3 1.0  
88 Audi A1 2016 14500 12100 30 56.5 1.4  
99 Audi A4 2017 15000 14864 145 52.3 1.4  
111 Audi A1 2016 10500 18272 0 67.3 1.0  
164 Audi A1 2017 11400 29515 0 67.3 1.0  
210 Audi Q2 2019 28985 824 145 40.9 1.5  
269 Audi A3 2018 19995 25000 150 42.2 1.5  
283 Audi Q3 2018 19485 7157 145 51.4 1.4

# Question 2: Exploratory Data Analysis (22 marks)

## Descriptive Statistics

1. What descriptive statistics would be appropriate for this dataset? Explain why these are useful in this context.

**(2 marks)**

### Answer:

The descriptive statistics that would be appropriate for this data set are as follows: 1. str() the structure function to perform basic inspection of the data frame 2. summary() the summary function to help perform a summary of the data frame either on each column or the entire data frame and returns a statistical information, which is helpful to check for central tendency (e.g mean, median) and spread. 3. describe() the describe function from psych help to show statistical description(central tendency) of the data frame.

1. Produce those descriptive statistics in the code chunk below:

**(4 marks)**

### Answer:

str(cars\_df)

'data.frame': 2000 obs. of 8 variables:  
 $ brand : chr "Audi" "Audi" "Audi" "Audi" ...  
 $ model : chr " A1" " A1" " A1" " A1" ...  
 $ year : int 2017 2016 2017 2016 2017 2016 2017 2019 2018 2018 ...  
 $ price : int 12500 11000 11700 14500 15000 10500 11400 28985 19995 19485 ...  
 $ mileage : int 15735 29946 19268 12100 14864 18272 29515 824 25000 7157 ...  
 $ tax : int 150 30 150 30 145 0 0 145 150 145 ...  
 $ mpg : num 55.4 55.4 67.3 56.5 52.3 67.3 67.3 40.9 42.2 51.4 ...  
 $ engineSize: num 1.4 1.4 1 1.4 1.4 1 1 1.5 1.5 1.4 ...

summary(cars\_df)

brand model year price   
 Length:2000 Length:2000 Min. :2000 Min. : 1294   
 Class :character Class :character 1st Qu.:2016 1st Qu.: 9554   
 Mode :character Mode :character Median :2017 Median :11897   
 Mean :2017 Mean :12981   
 3rd Qu.:2018 3rd Qu.:15700   
 Max. :2020 Max. :49999   
 mileage tax mpg engineSize   
 Min. : 1 Min. : 0.0 Min. :20.90 Min. :0.000   
 1st Qu.: 8780 1st Qu.:125.0 1st Qu.:51.40 1st Qu.:1.000   
 Median :15976 Median :145.0 Median :56.50 Median :1.000   
 Mean :18464 Mean :114.4 Mean :55.96 Mean :1.196   
 3rd Qu.:26147 3rd Qu.:145.0 3rd Qu.:62.80 3rd Qu.:1.400   
 Max. :59989 Max. :165.0 Max. :85.60 Max. :5.000

describe(cars\_df)

vars n mean sd median trimmed mad min max  
brand\* 1 2000 2.66 0.78 3.0 2.79 0.00 1.0 4.0  
model\* 2 2000 15.38 6.11 16.0 15.39 1.48 1.0 35.0  
year 3 2000 2017.25 1.67 2017.0 2017.44 1.48 2000.0 2020.0  
price 4 2000 12980.57 5027.20 11897.0 12434.44 4302.51 1294.0 49999.0  
mileage 5 2000 18464.31 12893.98 15975.5 17223.97 12436.05 1.0 59989.0  
tax 6 2000 114.44 54.74 145.0 123.88 0.00 0.0 165.0  
mpg 7 2000 55.96 8.05 56.5 56.73 7.56 20.9 85.6  
engineSize 8 2000 1.20 0.32 1.0 1.14 0.00 0.0 5.0  
 range skew kurtosis se  
brand\* 3.0 -1.36 0.66 0.02  
model\* 34.0 -0.05 1.10 0.14  
year 20.0 -1.44 6.46 0.04  
price 48705.0 1.33 3.22 112.41  
mileage 59988.0 0.82 0.22 288.32  
tax 165.0 -1.28 -0.22 1.22  
mpg 64.7 -0.71 0.43 0.18  
engineSize 5.0 2.78 21.76 0.01

1. What have those descriptive statistics told you – and how does this inform the analysis that you would undertake on this data or any additional data cleaning requirements?

**(4 marks)**

### Answer:

I have been able to see how the data frame is structured and also see the summary of the statistics performed on the data frame, the minimum value, maximum, mean, 25%, 75%, median, variance and standard deviation of the numerical variables. Also, from the statistical summary we can tell that mpg is symmetric due to the close value of mean and median. We can also see that tax is skewed due to the difference in mean and median.

## Exploratory Graphs

1. What exploratory graphs would be appropriate for this dataset? Explain why these are useful in this context.

**(2 marks)**

### Answer:

box plots: A very good way to show and identify the mean values, dispersion and skewness of the data set Histogram: A very good way to identify spread

1. Now produce those exploratory graphs in the code chunk below:

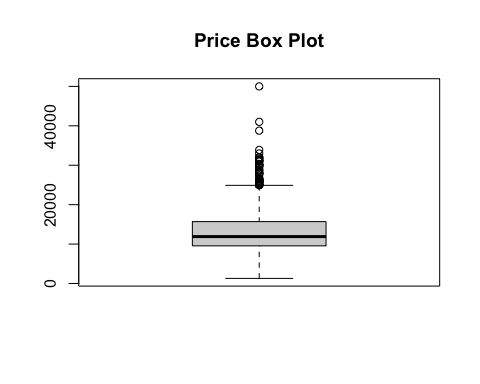
**(4 marks)**

### Answer:

# Using Box plots  
# for price  
summary(cars\_df$price)

Min. 1st Qu. Median Mean 3rd Qu. Max.   
 1294 9554 11897 12981 15700 49999

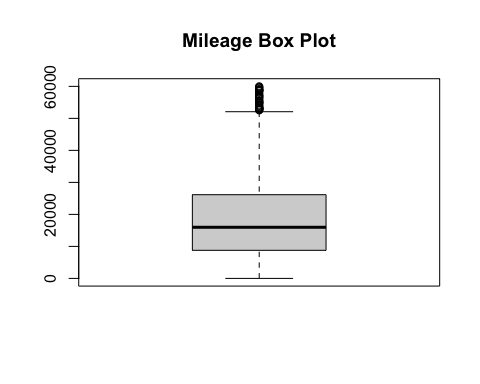
boxplot(cars\_df$price, main = "Price Box Plot")



# for mileage  
summary(cars\_df$mileage)

Min. 1st Qu. Median Mean 3rd Qu. Max.   
 1 8780 15976 18464 26147 59989

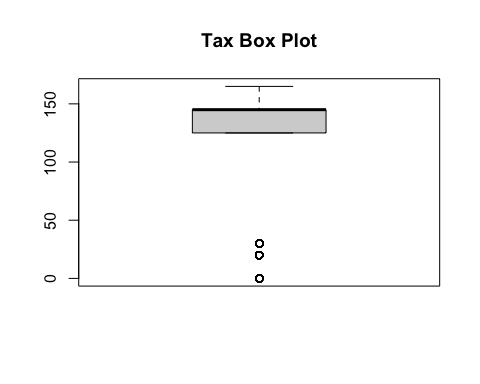
boxplot(cars\_df$mileage, main = "Mileage Box Plot")



# for tax   
summary(cars\_df$tax)

Min. 1st Qu. Median Mean 3rd Qu. Max.   
 0.0 125.0 145.0 114.4 145.0 165.0

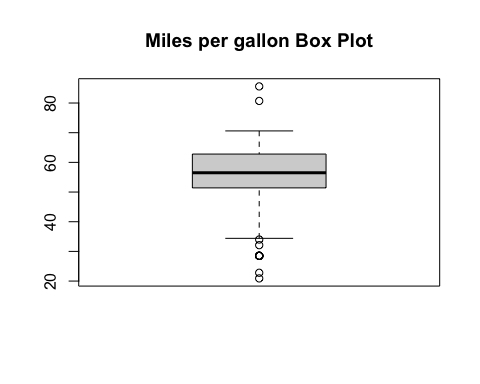
boxplot(cars\_df$tax, main = "Tax Box Plot")



# for mpg   
summary(cars\_df$mpg)

Min. 1st Qu. Median Mean 3rd Qu. Max.   
 20.90 51.40 56.50 55.96 62.80 85.60

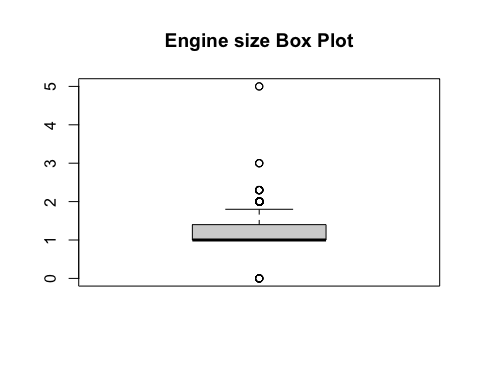
boxplot(cars\_df$mpg, main = "Miles per gallon Box Plot")



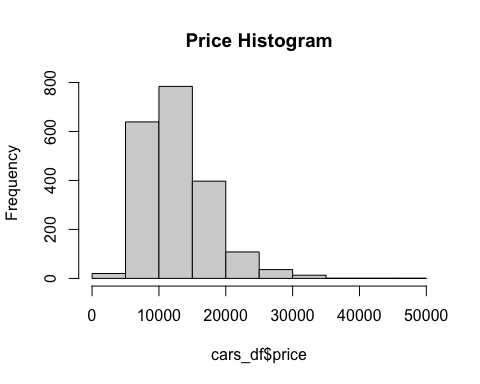
# for engine size  
summary(cars\_df$engineSize)

Min. 1st Qu. Median Mean 3rd Qu. Max.   
 0.000 1.000 1.000 1.196 1.400 5.000

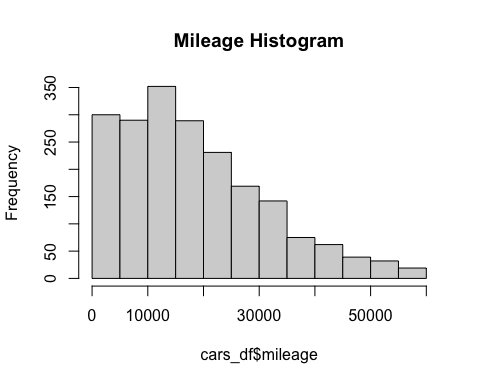
boxplot(cars\_df$engineSize, main = "Engine size Box Plot")



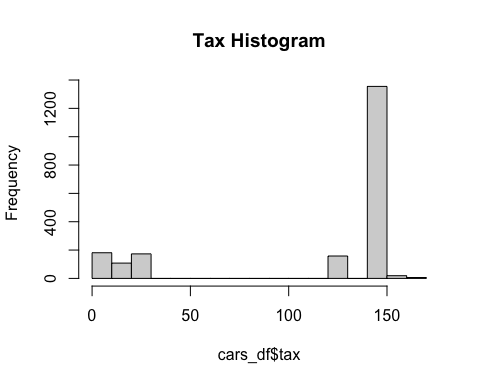
# Using Histogram for price and tax   
# price  
hist(cars\_df$price, main = "Price Histogram")



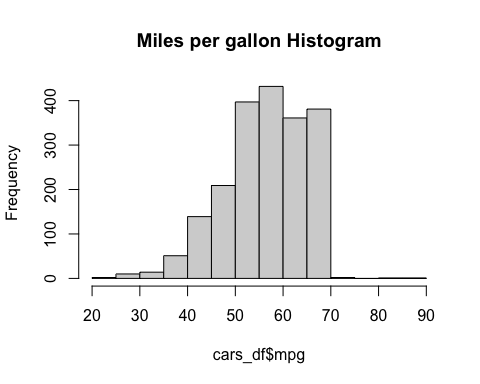
# Mileage  
hist(cars\_df$mileage, main = "Mileage Histogram")



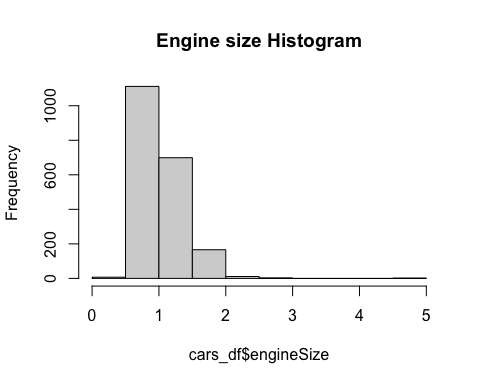
# tax  
hist(cars\_df$tax, main = "Tax Histogram")



# mpg   
hist(cars\_df$mpg, main = "Miles per gallon Histogram")



# engine size   
hist(cars\_df$engineSize, main = "Engine size Histogram")



1. Interpret these exploratory graphs. How do these graphs inform your subsequent analysis?

**(4 marks)**

### Answer:

Box plot of Price shows that it is positively skewed with outliers Box plot of Mileage shows that mileage is positive skewed too with outliers Box plot of Tax shows that it is negatively skewed with outliers Box plot of Miles per gallon shows that it is symmetric with outliers Box plot of Engine shows that it is positively skewed with outliers

Histogram for price: The graph is skewed right which shows that most of the price of the cars were between 10,000 to 20,000 and few prices between 30,000 and 35,000. Histogram for mileage: The graph is skewed right and shows the mileage is at the highest between 10,000 and 30,000

## Correlations

1. What linear correlations are present within this data?

**(2 marks)**

### Answer:

# Using Price  
cor(cars\_df$price, cars\_df$tax)

[1] 0.4098221

# 0.4098221 shows a weak positive linear relationship between price of a car and tax  
cor(cars\_df$price, cars\_df$year)

[1] 0.5970931

# 0.5970931 shows a moderate positive linear relationship between price of a car and year   
cor(cars\_df$price, cars\_df$mileage)

[1] -0.5380975

# -0.5380975 shows a moderate negative linear relationship between mileage and price of a car in the data set  
cor(cars\_df$price, cars\_df$engineSize)

[1] 0.5033365

# 0.5033365 shows a moderate positive linear relationship between price of a car and engine size   
cor(cars\_df$price, cars\_df$mpg)

[1] -0.6026653

# -0.6026653 shows a moderate negative linear relationship between miles per gallon and price of a car in the data set   
  
# Using year   
cor(cars\_df$year, cars\_df$tax)

[1] 0.5749106

# 0.5749106 shows moderate positive linear relationship between year and tax  
cor(cars\_df$year, cars\_df$mileage)

[1] -0.7046688

# -0.7046688 shows a strong negative linear relationship between year and mileage  
cor(cars\_df$year, cars\_df$engineSize)

[1] -0.03208153

# -0.03208153 shows almost no linear relation ship between year and engine size  
cor(cars\_df$year, cars\_df$mpg)

[1] -0.2067323

# -0.2067323 shows a weak negative linear relationship between year and mpg   
  
# Using tax   
cor(cars\_df$tax, cars\_df$mileage)

[1] -0.4605932

# -0.4605932 shows a moderate negative linear relationship between tax and mpg  
cor(cars\_df$tax, cars\_df$engineSize)

[1] 0.1920511

# 0.1920511 shows almost no linear relationship between tax and engine size  
cor(cars\_df$tax, cars\_df$mpg)

[1] -0.382534

# -0.382534 shows a weak negative linear relationship between tax and miles per gallon   
  
# using mileage   
cor(cars\_df$mileage, cars\_df$engineSize)

[1] -0.009642757

# -0.009642757 shows a very weak negative linear relationship between mileage and engine size  
cor(cars\_df$mileage, cars\_df$mpg)

[1] 0.2238517

# 0.2238517 shows a very weak positive linear relationship between mileage and miles per gallon  
  
# using miles per gallon  
cor(cars\_df$mpg, cars\_df$engineSize)

[1] -0.6770263

# -0.6770263 show a stong negative linear relationship between mpg and engine size

# Question 3: Bivariate relationship (14 marks)

1. Which of the potential explanatory variables has the strongest linear relationship with the dependent variable?

**(1 mark)**

### Answer:

From correlation performed above using the cor() function performed on Price(the explanatory variable) and year(dependent variable) had the strongest linear relationship at 0.5970931 which shows a moderate uphill positive linear relationship

1. Create a linear model to model this relationship.

**(2 marks)**

### Answer:

car\_model <- lm(price ~ year, cars\_df)  
summary(car\_model)

Call:  
lm(formula = price ~ year, data = cars\_df)  
  
Residuals:  
 Min 1Q Median 3Q Max   
 -9214 -2762 -914 1930 37470   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) -3.603e+06 1.087e+05 -33.15 <2e-16 \*\*\*  
year 1.793e+03 5.388e+01 33.27 <2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 4034 on 1998 degrees of freedom  
Multiple R-squared: 0.3565, Adjusted R-squared: 0.3562   
F-statistic: 1107 on 1 and 1998 DF, p-value: < 2.2e-16

1. Explain and interpret the model:

**(3 marks)**

### Answer:

The formula call shown in the summary returned is the formula used to fit the data. The Residual of the model which is the difference between observed response values and the value the model predicted which is broken down into five summary points min value of -9214, 1Q -2762 , median -914, 3Q 1930 and max 37,470. The model strongly predicts that certain points that falls from the actual observed points. The coefficient estimate shows the intercept, The p-value is considerably close to zero and it shows significance. The regression model explains 35.65% of the variation on the response. That is what the R-squared shows us, Which is not so high but fair enough considering the model has a lot of other variables, maybe if we add these variables, the variation might increase. the analysis also shows, just because the explanatory variable is highly correlated with the dependent variable doesn’t mean the model will be perfect The model equation is Price = -3.603e+06 + 1.793e+03(year)

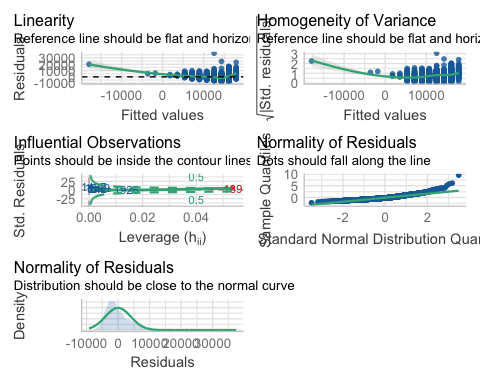
1. Comment on the performance of this model, including comments on overall model fit and the validity of model assumptions. Include any additional code required for you to make these comments in the code chunk below.

**(4 marks)**

### Answer:

Plotting the residuals, the observations are independent according to plot from the check\_model function The fitted plot against residuals also show that there is constant spread. The plot shows the data is distributed normally and there are outliers in the data Checking Linearity between the residuals and covariates, we want to have a scatter plot where each value is above or below the zero line. We do not see any great pattern with any of the covariates We check for goodness of fit between the predicted values and the observed values for the response variable. It seems the values are mostly fitted around the middle and show a linear relationship

check\_model(car\_model)



## Bootstrap

1. Use bootstrapping on this model to obtain a 95% confidence interval of the estimate of the slope parameter.

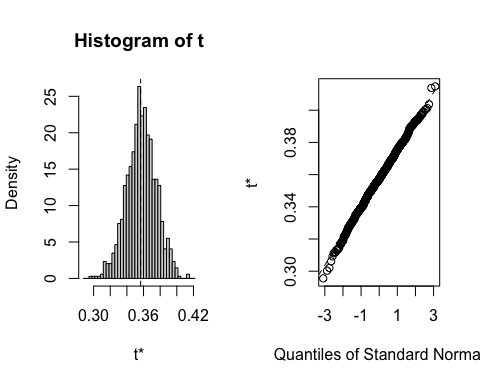
**(4 marks)**

### Answer:

# Bootstrap 95% CI for R-Squared  
# function to obtain R-Squared from the data   
rsq <- function(formula, data, indices) {  
 d <- data[indices,] # allows boot to select sample   
 fit <- lm(formula, data = d)  
 return(summary(fit)$r.square)  
}   
# bootstrapping with 1000 replications   
obs <- boot(data = cars\_df, statistic= rsq,   
 R = 1000, formula = price ~ year)  
  
# view observation  
obs

ORDINARY NONPARAMETRIC BOOTSTRAP  
  
  
Call:  
boot(data = cars\_df, statistic = rsq, R = 1000, formula = price ~   
 year)  
  
  
Bootstrap Statistics :  
 original bias std. error  
t1\* 0.3565202 0.0008127726 0.01784249

plot(obs)



# get 95% confidence interval   
boot.ci(results, type = c("norm", "basic", "perc", "stud"),   
 h = log, hdot = function(x) 1/x)

Error in boot.ci(results, type = c("norm", "basic", "perc", "stud"), h = log, : object 'results' not found

# Question 4: Multivariable relationship (10 marks)

Create a model with all of the appropriate remaining explanatory variables included:

car\_model2 <- lm(price ~ year + mileage + tax + mpg + engineSize, data = cars\_df)  
summary(car\_model2)

Call:  
lm(formula = price ~ year + mileage + tax + mpg + engineSize,   
 data = cars\_df)  
  
Residuals:  
 Min 1Q Median 3Q Max   
 -8577 -1484 -67 1409 16407   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) -3.025e+06 1.185e+05 -25.535 < 2e-16 \*\*\*  
year 1.508e+03 5.867e+01 25.707 < 2e-16 \*\*\*  
mileage -7.170e-02 6.964e-03 -10.295 < 2e-16 \*\*\*  
tax -1.229e+01 1.497e+00 -8.209 3.96e-16 \*\*\*  
mpg -1.624e+02 1.153e+01 -14.090 < 2e-16 \*\*\*  
engineSize 5.749e+03 2.756e+02 20.862 < 2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 2822 on 1994 degrees of freedom  
Multiple R-squared: 0.6858, Adjusted R-squared: 0.685   
F-statistic: 870.3 on 5 and 1994 DF, p-value: < 2.2e-16

1. Explain and interpret the model:

**(4 marks)**

### Answer:

It can be said that p-value of the F-statistic is <2.2e-16 which is highly significant. It can be said that at least one of the predictor variable is significantly related to the outcome variable. The model equation is price = -3.025e+06 + 1.508e+03(year) + -7.170e-02(mileage) + -1.229e+01(tax) + -1.624e+02(mpg) + 5.749e+03(engineSize)

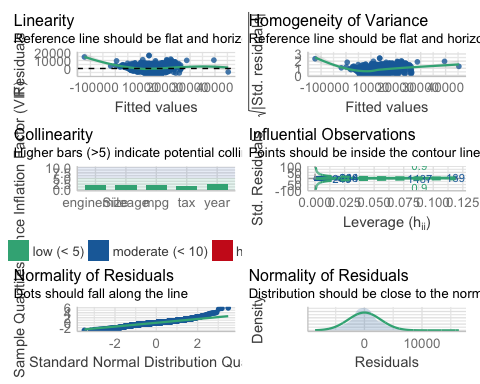
1. Comment on the performance of this model, including comments on overall model fit and the validity of model assumptions. Include any additional code required for you to make these comments in the code chunk below.

**(4 marks)**

### Answer:

Plotting the residuals, the observations are independent according to plot from the check\_model function The fitted plot against residuals also show that there is constant spread. The plot shows the data is distributed normally and there are outliers in the data Checking Linearity between the residuals and covariates, we want to have a scatter plot where each value is above or below the zero line. We do not see any great pattern with any of the covariates We check for goodness of fit between the predicted values and the observed values for the response variable. It seems the values are mostly fitted around the middle and show a linear relationship

check\_model(car\_model2)



1. What general concerns do you have regarding this model?

**(2 marks)**

### Answer:

Data must be independent Sensitivity to outliers because outliers can cause a huge effect on the regression model. The model only looks at the mean of the dependent variables

# Question 5: Model simplification (8 marks)

1. What approaches for model simplification would you consider implementing and why?

**(4 marks)**

### Answer:

ANOVA drop1 function() AIC method; it makes the model usually better by picking only the explanatory variables that are significant to the dependent variables and using that to form your model

1. What are the potential advantages of simplifying a model?

**(2 marks)**

### Answer:

A potential advantage in simplifying a model is that it makes it work better. It removes a non-significant explanatory variables.

1. What are the potential disadvantages of simplifying a model?

**(2 marks)**

### Answer:

Model simplification is not as straightforward as removing or dropping explanatory variables. A potential disadvantage is that the more you remove the explanatory variables, the weaker the model as it reduces the R-squared value which explains the variation of the model

# Question 6: Reporting (35 marks)

A client is looking to purchase a used Skoda Superb (registration year either 2018 or 2019, manual transmission, petrol engine) and wants to understand what factors influence the expected price of a used car, (and how they influence the price).

Write a short report of 300-500 words for the client.

Furthermore, include an explanation as to which model you would recommend, and why you have selected that model.

Comment on any suggestions for alterations to the model that would be appropriate to consider.

Highlight what may or may not be directly transferable from the scenario analysed in Questions 1 to 5.

### Answer:

# Report

This data set aim is to to find and understand what factors influence the expected price of a used car and how they influence the price, and whether other variables that might affect or influence the price of a used car . We can also check if the mileage of a used car can affect the price. This is particularly important in this study because we may potentially use our findings on other data sets too.

I have been able to establish relationships within variable of the data set that affects the price of a used car and that which has little or no affect on the price.

Price vs Year of car registration; Analysis performed on the data set shows that the higher the year of registration of a used car, the higher the amount of the car. Thus, I have been able to establish that the higher the year of registration of a used car the costlier the used car.

Price vs Manual Transmission; This was filtered out from the analysis because it does not affect the price of a used car. Thus, I have been able to establish that the manual transmission of a used car does not influence the price a used car.

Price vs Petrol engine; Like manual transmission, petrol engine also do not affect the price of a used car. It was filtered out from the data set analysis was made. Thus, I was able to establish that petrol engine a used car run on does not influence on the price of a used car.

Price vs Engine Size; on the other hand, engine size unlike petrol engine have affect the price of a used car. The larger the size of a used car the higher the price of the car. Thus I was able to establish that size of a car engine have moderate influence on the price of a used car.

Price Vs Tax; tax charge on a used car have a very weak effect on the price of a used car. Thus, I was able to establish that tax have little influence on the price of a used car.

Price vs Miles per gallon; miles per gallon of a used car do not affect the price of a used car. Thus I was able to establish that miles per gallon do not influence the price of a used car.

Price vs Mileage; mileage of a used car do not affect the price of a used car. Thus I was able to establish that mileage of a used car do not influence the price of a used car.

# Recommendation

I’ll recommend a Ford Fiesta year 2019 at a price of 14,995. I’ll recommend a Ford Fiesta because the year of registration falls between the year of choice the client wants his preferred car. also, there share similarities in fuel supply, engine position and the car has a low mileage of 15 even though mileage has no influence on price of the car.

# Session Information

Do not edit this part. Make sure that you compile your document so that the information about your session (including software / package versions) is included in your submission.

sessionInfo()

R version 4.1.1 (2021-08-10)  
Platform: x86\_64-apple-darwin17.0 (64-bit)  
Running under: macOS Catalina 10.15.7  
  
Matrix products: default  
BLAS: /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRblas.0.dylib  
LAPACK: /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRlapack.dylib  
  
locale:  
[1] en\_US.UTF-8/en\_US.UTF-8/en\_US.UTF-8/C/en\_US.UTF-8/en\_US.UTF-8  
  
attached base packages:  
[1] stats graphics grDevices utils datasets methods base   
  
other attached packages:  
 [1] boot\_1.3-28 patchwork\_1.1.1 see\_0.6.8 performance\_0.8.0  
 [5] psych\_2.1.9 forcats\_0.5.1 stringr\_1.4.0 dplyr\_1.0.7   
 [9] purrr\_0.3.4 readr\_2.0.2 tidyr\_1.1.4 tibble\_3.1.5   
[13] ggplot2\_3.3.5 tidyverse\_1.3.1   
  
loaded via a namespace (and not attached):  
 [1] Rcpp\_1.0.7 lubridate\_1.8.0 lattice\_0.20-44 assertthat\_0.2.1   
 [5] digest\_0.6.28 utf8\_1.2.2 R6\_2.5.1 cellranger\_1.1.0   
 [9] backports\_1.3.0 reprex\_2.0.1 evaluate\_0.14 highr\_0.9   
[13] httr\_1.4.2 pillar\_1.6.3 rlang\_0.4.11 readxl\_1.3.1   
[17] rstudioapi\_0.13 Matrix\_1.3-4 rmarkdown\_2.11 labeling\_0.4.2   
[21] splines\_4.1.1 munsell\_0.5.0 broom\_0.7.10 compiler\_4.1.1   
[25] modelr\_0.1.8 xfun\_0.26 pkgconfig\_2.0.3 mnormt\_2.0.2   
[29] tmvnsim\_1.0-2 mgcv\_1.8-36 htmltools\_0.5.2 insight\_0.15.0   
[33] tidyselect\_1.1.1 fansi\_0.5.0 crayon\_1.4.1 tzdb\_0.1.2   
[37] dbplyr\_2.1.1 withr\_2.4.2 grid\_4.1.1 nlme\_3.1-152   
[41] jsonlite\_1.7.2 gtable\_0.3.0 lifecycle\_1.0.1 DBI\_1.1.1   
[45] bayestestR\_0.11.5 magrittr\_2.0.1 scales\_1.1.1 datawizard\_0.2.2   
[49] cli\_3.0.1 stringi\_1.7.5 farver\_2.1.0 fs\_1.5.0   
[53] xml2\_1.3.2 ellipsis\_0.3.2 generics\_0.1.0 vctrs\_0.3.8   
[57] tools\_4.1.1 glue\_1.4.2 hms\_1.1.1 parallel\_4.1.1   
[61] fastmap\_1.1.0 yaml\_2.2.1 colorspace\_2.0-2 rvest\_1.0.2   
[65] knitr\_1.36 haven\_2.4.3