Quantium_Data_Analytics_Job_Simulation

Divine Gatebuka Iradukunda

2025-10-22

Opening the Transaction Data file

	DA… <int>	STORE_… <int>	LYLTY_CARD_… <int>	TXN… <int>	PROD… <int>	PROD_NAME <chr>
1	43390	1	1000	1	5	Natural Chip Compny SeaSalt175g
2	43599	1	1307	348	66	CCs Nacho Cheese 175g
3	43605	1	1343	383	61	Smiths Crinkle Cut Chips Chicken 170g
4	43329	2	2373	974	69	Smiths Chip Thinly S/Cream&Onion 175g
5	43330	2	2426	1038	108	Kettle Tortilla ChpsHny&Jlpno Chili 150g
6	43604	4	4074	2982	57	Old El Paso Salsa Dip Tomato Mild 300g

6 rows | 1-7 of 9 columns

Opening the Purchase Behaviour file

	LYLTY_CARD_NBR <int>	LIFESTAGE <chr>	PREMIUM_CUSTOMER <chr>
1	1000	YOUNG SINGLES/COUPLES	Premium
2	1002	YOUNG SINGLES/COUPLES	Mainstream
3	1003	YOUNG FAMILIES	Budget
4	1004	OLDER SINGLES/COUPLES	Mainstream
5	1005	MIDAGE SINGLES/COUPLES	Mainstream
6	1007	YOUNG SINGLES/COUPLES	Budget

6 rows

Exploratory Data Analysis

	DATE <date>	STORE_… <int>	LYLTY_CARD_… <int>	TXN… <int>	PROD… <int>	PROD_NAME <chr>
	2018-10-17	1	1000	1	5	Natural Chip Compny SeaSalt175g

DATE <date>	STORE_… <int>	LYLTY_CARD_… <int>	TXN… <int>	PROD… <int>	PROD_NAME <chr>
2019-05-14	1	1307	348	66	CCs Nacho Cheese 175g
2019-05-20	1	1343	383	61	Smiths Crinkle Cut Chips Chicken 170g
2018-08-17	2	2373	974	69	Smiths Chip Thinly S/Cream&Onion 175g
2018-08-18	2	2426	1038	108	Kettle Tortilla ChpsHny&Jlpno Chili 150g
2019-05-19	4	4074	2982	57	Old El Paso Salsa Dip Tomato Mild 300g

Examining PROD_NAME

```
##   [1] "Natural Chip        Compny SeaSalt175g"
##   [2] "CCs Nacho Cheese    175g"
##   [3] "Smiths Crinkle Cut  Chips Chicken 170g"
##   [4] "Smiths Chip Thinly  S/Cream&Onion 175g"
##   [5] "Kettle Tortilla ChpsHny&Jlpno Chili 150g"
##   [6] "Old El Paso Salsa   Dip Tomato Mild 300g"
##   [7] "Smiths Crinkle Chips Salt & Vinegar 330g"
##   [8] "Grain Waves         Sweet Chilli 210g"
##   [9] "Doritos Corn Chip Mexican Jalapeno 150g"
##  [10] "Grain Waves Sour    Cream&Chives 210G"
##  [11] "Kettle Sensations   Siracha Lime 150g"
##  [12] "Twisties Cheese     270g"
##  [13] "WW Crinkle Cut      Chicken 175g"
##  [14] "Thins Chips Light&  Tangy 175g"
##  [15] "CCs Original 175g"
##  [16] "Burger Rings 220g"
##  [17] "NCC Sour Cream &    Garden Chives 175g"
##  [18] "Doritos Corn Chip Southern Chicken 150g"
##  [19] "Cheezels Cheese Box 125g"
##  [20] "Smiths Crinkle      Original 330g"
##  [21] "Infzns Crn Crnchers Tangy Gcamole 110g"
##  [22] "Kettle Sea Salt     And Vinegar 175g"
##  [23] "Smiths Chip Thinly  Cut Original 175g"
##  [24] "Kettle Original 175g"
##  [25] "Red Rock Deli Thai  Chilli&Lime 150g"
##  [26] "Pringles Sthrn FriedChicken 134g"
##  [27] "Pringles Sweet&Spcy BBQ 134g"
##  [28] "Red Rock Deli SR    Salsa & Mzzrlla 150g"
##  [29] "Thins Chips         Originl saltd 175g"
##  [30] "Red Rock Deli Sp    Salt & Truffle 150G"
##  [31] "Smiths Thinly       Swt Chli&S/Cream175G"
##  [32] "Kettle Chilli 175g"
##  [33] "Doritos Mexicana    170g"
##  [34] "Smiths Crinkle Cut  French OnionDip 150g"
##  [35] "Natural ChipCo      Hony Soy Chckn175g"
##  [36] "Dorito Corn Chp     Supreme 380g"
##  [37] "Twisties Chicken270g"
##  [38] "Smiths Thinly Cut   Roast Chicken 175g"
##  [39] "Smiths Crinkle Cut  Tomato Salsa 150g"
##  [40] "Kettle Mozzarella   Basil & Pesto 175g"
##  [41] "Infuzions Thai SweetChili PotatoMix 110g"
##  [42] "Kettle Sensations   Camembert & Fig 150g"
##  [43] "Smith Crinkle Cut   Mac N Cheese 150g"
##  [44] "Kettle Honey Soy    Chicken 175g"
##  [45] "Thins Chips Seasonedchicken 175g"
##  [46] "Smiths Crinkle Cut  Salt & Vinegar 170g"
##  [47] "Infuzions BBQ Rib   Prawn Crackers 110g"
##  [48] "GrnWves Plus Btroot & Chilli Jam 180g"
##  [49] "Tyrrells Crisps     Lightly Salted 165g"
##  [50] "Kettle Sweet Chilli And Sour Cream 175g"
##  [51] "Doritos Salsa       Medium 300g"
##  [52] "Kettle 135g Swt Pot Sea Salt"
```

```
##  [53] "Pringles SourCream  Onion 134g"
##  [54] "Doritos Corn Chips  Original 170g"
##  [55] "Twisties Cheese     Burger 250g"
##  [56] "Old El Paso Salsa   Dip Chnky Tom Ht300g"
##  [57] "Cobs Popd Swt/Chlli &Sr/Cream Chips 110g"
##  [58] "Woolworths Mild     Salsa 300g"
##  [59] "Natural Chip Co     Tmato Hrb&Spce 175g"
##  [60] "Smiths Crinkle Cut  Chips Original 170g"
##  [61] "Cobs Popd Sea Salt  Chips 110g"
##  [62] "Smiths Crinkle Cut  Chips Chs&Onion170g"
##  [63] "French Fries Potato Chips 175g"
##  [64] "Old El Paso Salsa   Dip Tomato Med 300g"
##  [65] "Doritos Corn Chips  Cheese Supreme 170g"
##  [66] "Pringles Original   Crisps 134g"
##  [67] "RRD Chilli&         Coconut 150g"
##  [68] "WW Original Corn    Chips 200g"
##  [69] "Thins Potato Chips  Hot & Spicy 175g"
##  [70] "Cobs Popd Sour Crm  &Chives Chips 110g"
##  [71] "Smiths Crnkle Chip  Orgnl Big Bag 380g"
##  [72] "Doritos Corn Chips  Nacho Cheese 170g"
##  [73] "Kettle Sensations   BBQ&Maple 150g"
##  [74] "WW D/Style Chip     Sea Salt 200g"
##  [75] "Pringles Chicken    Salt Crips 134g"
##  [76] "WW Original Stacked Chips 160g"
##  [77] "Smiths Chip Thinly  CutSalt/Vinegr175g"
##  [78] "Cheezels Cheese 330g"
##  [79] "Tostitos Lightly    Salted 175g"
##  [80] "Thins Chips Salt &  Vinegar 175g"
##  [81] "Smiths Crinkle Cut  Chips Barbecue 170g"
##  [82] "Cheetos Puffs 165g"
##  [83] "RRD Sweet Chilli &  Sour Cream 165g"
##  [84] "WW Crinkle Cut      Original 175g"
##  [85] "Tostitos Splash Of  Lime 175g"
##  [86] "Woolworths Medium   Salsa 300g"
##  [87] "Kettle Tortilla ChpsBtroot&Ricotta 150g"
##  [88] "CCs Tasty Cheese    175g"
##  [89] "Woolworths Cheese   Rings 190g"
##  [90] "Tostitos Smoked     Chipotle 175g"
##  [91] "Pringles Barbeque   134g"
##  [92] "WW Supreme Cheese   Corn Chips 200g"
##  [93] "Pringles Mystery    Flavour 134g"
##  [94] "Tyrrells Crisps     Ched & Chives 165g"
##  [95] "Snbts Whlgrn Crisps Cheddr&Mstrd 90g"
##  [96] "Cheetos Chs & Bacon Balls 190g"
##  [97] "Pringles Slt Vingar 134g"
##  [98] "Infuzions SourCream&Herbs Veg Strws 110g"
##  [99] "Kettle Tortilla ChpsFeta&Garlic 150g"
## [100] "Infuzions Mango     Chutny Papadums 70g"
## [101] "RRD Steak &         Chimuchurri 150g"
## [102] "RRD Honey Soy       Chicken 165g"
## [103] "Sunbites Whlegrn    Crisps Frch/Onin 90g"
## [104] "RRD Salt & Vinegar  165g"
```

```
## [105] "Doritos Cheese      Supreme 330g"
## [106] "Smiths Crinkle Cut  Snag&Sauce 150g"
## [107] "WW Sour Cream &OnionStacked Chips 160g"
## [108] "RRD Lime & Pepper   165g"
## [109] "Natural ChipCo Sea  Salt & Vinegr 175g"
## [110] "Red Rock Deli Chikn&Garlic Aioli 150g"
## [111] "RRD SR Slow Rst     Pork Belly 150g"
## [112] "RRD Pc Sea Salt     165g"
## [113] "Smith Crinkle Cut   Bolognese 150g"
## [114] "Doritos Salsa Mild  300g"
```

Basic Text Analysis

words <chr>
Natural
Chip
Compny
SeaSalt175g
CCs
Nacho

6 rows

```
## [1] 589
```

```
## [1] "Natural"     "Chip"        "Compny"      "SeaSalt175g" "CCs"
## [6] "Nacho"
```

```
## [1] "Bolognese" "150g"      "Doritos"   "Salsa"     "Mild"      "300g"
```

```
## [1] 589
```

Remove digits and special characters

```
## [1] TRUE
```

```
## [1] 37
```

```
## [1] "there are 37 words in the productWords table that contain &, now remove that and every o
ther special character"
```

Observation: Now we know that there are 37 words in the productWords table that contain "&", now let us find all special characters.

```
##   [1] "Natural"          "Chip"             "Compny"
##   [4] "SeaSaltg"         "CCs"              "Nacho"
##   [7] "Cheese"           "g"                "Smiths"
##  [10] "Crinkle"          "Cut"              "Chips"
##  [13] "Chicken"          "g"                "Smiths"
##  [16] "Chip"             "Thinly"           "SCreamOnion"
##  [19] "g"                "Kettle"           "Tortilla"
##  [22] "ChpsHnyJlpno"     "Chili"            "g"
##  [25] "Old"              "El"               "Paso"
##  [28] "Salsa"            "Dip"              "Tomato"
##  [31] "Mild"             "g"                "Smiths"
##  [34] "Crinkle"          "Chips"            "Salt"
##  [37] ""                 "Vinegar"          "g"
##  [40] "Grain"            "Waves"            "Sweet"
##  [43] "Chilli"           "g"                "Doritos"
##  [46] "Corn"             "Chip"             "Mexican"
##  [49] "Jalapeno"         "g"                "Grain"
##  [52] "Waves"            "Sour"             "CreamChives"
##  [55] "G"                "Kettle"           "Sensations"
##  [58] "Siracha"          "Lime"             "g"
##  [61] "Twisties"         "Cheese"           "g"
##  [64] "WW"               "Crinkle"          "Cut"
##  [67] "Chicken"          "g"                "Thins"
##  [70] "Chips"            "Light"            "Tangy"
##  [73] "g"                "CCs"              "Original"
##  [76] "g"                "Burger"           "Rings"
##  [79] "g"                "NCC"              "Sour"
##  [82] "Cream"            ""                 "Garden"
##  [85] "Chives"           "g"                "Doritos"
##  [88] "Corn"             "Chip"             "Southern"
##  [91] "Chicken"          "g"                "Cheezels"
##  [94] "Cheese"           "Box"              "g"
##  [97] "Smiths"           "Crinkle"          "Original"
## [100] "g"                "Infzns"           "Crn"
## [103] "Crnchers"         "Tangy"            "Gcamole"
## [106] "g"                "Kettle"           "Sea"
## [109] "Salt"             "And"              "Vinegar"
## [112] "g"                "Smiths"           "Chip"
## [115] "Thinly"           "Cut"              "Original"
## [118] "g"                "Kettle"           "Original"
## [121] "g"                "Red"              "Rock"
## [124] "Deli"             "Thai"             "ChilliLime"
## [127] "g"                "Pringles"         "Sthrn"
## [130] "FriedChicken"     "g"                "Pringles"
## [133] "SweetSpcy"        "BBQ"              "g"
## [136] "Red"              "Rock"             "Deli"
## [139] "SR"               "Salsa"            ""
## [142] "Mzzrlla"          "g"                "Thins"
## [145] "Chips"            "Originl"          "saltd"
## [148] "g"                "Red"              "Rock"
## [151] "Deli"             "Sp"               "Salt"
## [154] ""                 "Truffle"          "G"
```

```
## [157] "Smiths"           "Thinly"           "Swt"
## [160] "ChliSCreamG"      "Kettle"           "Chilli"
## [163] "g"                "Doritos"          "Mexicana"
## [166] "g"                "Smiths"           "Crinkle"
## [169] "Cut"              "French"           "OnionDip"
## [172] "g"                "Natural"          "ChipCo"
## [175] "Hony"             "Soy"              "Chckng"
## [178] "Dorito"           "Corn"             "Chp"
## [181] "Supreme"          "g"                "Twisties"
## [184] "Chickeng"         "Smiths"           "Thinly"
## [187] "Cut"              "Roast"            "Chicken"
## [190] "g"                "Smiths"           "Crinkle"
## [193] "Cut"              "Tomato"           "Salsa"
## [196] "g"                "Kettle"           "Mozzarella"
## [199] "Basil"            ""                 "Pesto"
## [202] "g"                "Infuzions"        "Thai"
## [205] "SweetChili"       "PotatoMix"        "g"
## [208] "Kettle"           "Sensations"       "Camembert"
## [211] ""                 "Fig"              "g"
## [214] "Smith"            "Crinkle"          "Cut"
## [217] "Mac"              "N"                "Cheese"
## [220] "g"                "Kettle"           "Honey"
## [223] "Soy"              "Chicken"          "g"
## [226] "Thins"            "Chips"            "Seasonedchicken"
## [229] "g"                "Smiths"           "Crinkle"
## [232] "Cut"              "Salt"             ""
## [235] "Vinegar"          "g"                "Infuzions"
## [238] "BBQ"              "Rib"              "Prawn"
## [241] "Crackers"         "g"                "GrnWves"
## [244] "Plus"             "Btroot"           ""
## [247] "Chilli"           "Jam"              "g"
## [250] "Tyrrells"         "Crisps"           "Lightly"
## [253] "Salted"           "g"                "Kettle"
## [256] "Sweet"            "Chilli"           "And"
## [259] "Sour"             "Cream"            "g"
## [262] "Doritos"          "Salsa"            "Medium"
## [265] "g"                "Kettle"           "g"
## [268] "Swt"              "Pot"              "Sea"
## [271] "Salt"             "Pringles"         "SourCream"
## [274] "Onion"            "g"                "Doritos"
## [277] "Corn"             "Chips"            "Original"
## [280] "g"                "Twisties"         "Cheese"
## [283] "Burger"           "g"                "Old"
## [286] "El"               "Paso"             "Salsa"
## [289] "Dip"              "Chnky"            "Tom"
## [292] "Htg"              "Cobs"             "Popd"
## [295] "SwtChlli"         "SrCream"          "Chips"
## [298] "g"                "Woolworths"       "Mild"
## [301] "Salsa"            "g"                "Natural"
## [304] "Chip"             "Co"               "Tmato"
## [307] "HrbSpce"          "g"                "Smiths"
## [310] "Crinkle"          "Cut"              "Chips"
```

```
## [313] "Original"          "g"               "Cobs"
## [316] "Popd"              "Sea"             "Salt"
## [319] "Chips"             "g"               "Smiths"
## [322] "Crinkle"           "Cut"             "Chips"
## [325] "ChsOniong"         "French"          "Fries"
## [328] "Potato"            "Chips"           "g"
## [331] "Old"               "El"              "Paso"
## [334] "Salsa"             "Dip"             "Tomato"
## [337] "Med"               "g"               "Doritos"
## [340] "Corn"              "Chips"           "Cheese"
## [343] "Supreme"           "g"               "Pringles"
## [346] "Original"          "Crisps"          "g"
## [349] "RRD"               "Chilli"          "Coconut"
## [352] "g"                 "WW"              "Original"
## [355] "Corn"              "Chips"           "g"
## [358] "Thins"             "Potato"          "Chips"
## [361] "Hot"               ""                "Spicy"
## [364] "g"                 "Cobs"            "Popd"
## [367] "Sour"              "Crm"             "Chives"
## [370] "Chips"             "g"               "Smiths"
## [373] "Crnkle"            "Chip"            "Orgnl"
## [376] "Big"               "Bag"             "g"
## [379] "Doritos"           "Corn"            "Chips"
## [382] "Nacho"             "Cheese"          "g"
## [385] "Kettle"            "Sensations"      "BBQMaple"
## [388] "g"                 "WW"              "DStyle"
## [391] "Chip"              "Sea"             "Salt"
## [394] "g"                 "Pringles"        "Chicken"
## [397] "Salt"              "Crips"           "g"
## [400] "WW"                "Original"        "Stacked"
## [403] "Chips"             "g"               "Smiths"
## [406] "Chip"              "Thinly"          "CutSaltVinegrg"
## [409] "Cheezels"          "Cheese"          "g"
## [412] "Tostitos"          "Lightly"         "Salted"
## [415] "g"                 "Thins"           "Chips"
## [418] "Salt"              ""                "Vinegar"
## [421] "g"                 "Smiths"          "Crinkle"
## [424] "Cut"               "Chips"           "Barbecue"
## [427] "g"                 "Cheetos"         "Puffs"
## [430] "g"                 "RRD"             "Sweet"
## [433] "Chilli"            ""                "Sour"
## [436] "Cream"             "g"               "WW"
## [439] "Crinkle"           "Cut"             "Original"
## [442] "g"                 "Tostitos"        "Splash"
## [445] "Of"                "Lime"            "g"
## [448] "Woolworths"        "Medium"          "Salsa"
## [451] "g"                 "Kettle"          "Tortilla"
## [454] "ChpsBtrootRicotta" "g"               "CCs"
## [457] "Tasty"             "Cheese"          "g"
## [460] "Woolworths"        "Cheese"          "Rings"
## [463] "g"                 "Tostitos"        "Smoked"
## [466] "Chipotle"          "g"               "Pringles"
```

```
## [469] "Barbeque"          "g"                "WW"
## [472] "Supreme"           "Cheese"           "Corn"
## [475] "Chips"             "g"                "Pringles"
## [478] "Mystery"           "Flavour"          "g"
## [481] "Tyrrells"          "Crisps"           "Ched"
## [484] ""                  "Chives"           "g"
## [487] "Snbts"             "Whlgrn"           "Crisps"
## [490] "CheddrMstrd"       "g"                "Cheetos"
## [493] "Chs"               ""                 "Bacon"
## [496] "Balls"             "g"                "Pringles"
## [499] "Slt"               "Vingar"           "g"
## [502] "Infuzions"         "SourCreamHerbs"   "Veg"
## [505] "Strws"             "g"                "Kettle"
## [508] "Tortilla"          "ChpsFetaGarlic"   "g"
## [511] "Infuzions"         "Mango"            "Chutny"
## [514] "Papadums"          "g"                "RRD"
## [517] "Steak"             ""                 "Chimuchurri"
## [520] "g"                 "RRD"              "Honey"
## [523] "Soy"               "Chicken"          "g"
## [526] "Sunbites"          "Whlegrn"          "Crisps"
## [529] "FrchOnin"          "g"                "RRD"
## [532] "Salt"              ""                 "Vinegar"
## [535] "g"                 "Doritos"          "Cheese"
## [538] "Supreme"           "g"                "Smiths"
## [541] "Crinkle"           "Cut"              "SnagSauce"
## [544] "g"                 "WW"               "Sour"
## [547] "Cream"             "OnionStacked"     "Chips"
## [550] "g"                 "RRD"              "Lime"
## [553] ""                  "Pepper"           "g"
## [556] "Natural"           "ChipCo"           "Sea"
## [559] "Salt"              ""                 "Vinegr"
## [562] "g"                 "Red"              "Rock"
## [565] "Deli"              "ChiknGarlic"      "Aioli"
## [568] "g"                 "RRD"              "SR"
## [571] "Slow"              "Rst"              "Pork"
## [574] "Belly"             "g"                "RRD"
## [577] "Pc"                "Sea"              "Salt"
## [580] "g"                 "Smith"            "Crinkle"
## [583] "Cut"               "Bolognese"        "g"
## [586] "Doritos"           "Salsa"            "Mild"
## [589] "g"
```

Sorting distinct words by frequency of occurence

clean	N
<chr>	<int>
g	105

clean <chr>	N <int>
Chips	21
Smiths	16
Crinkle	14
Cut	14
Kettle	13
Cheese	12
Salt	12
Original	10
Chip	9

1-10 of 196 rows Previous **1** 2 3 4 5 6 … 20 Next

Remove Salsa products

```
## [1] "Salsa product count after filtering: 0"
```

Summary Statistics

```
##       DATE              STORE_NBR     LYLTY_CARD_NBR        TXN_ID
##  Min.   :2018-07-01   Min.   :  1.0   Min.   :   1000   Min.   :      1
##  1st Qu.:2018-09-30   1st Qu.: 70.0   1st Qu.:  70015   1st Qu.:  67569
##  Median :2018-12-30   Median :130.0   Median : 130367   Median : 135183
##  Mean   :2018-12-30   Mean   :135.1   Mean   : 135531   Mean   : 135131
##  3rd Qu.:2019-03-31   3rd Qu.:203.0   3rd Qu.: 203084   3rd Qu.: 202654
##  Max.   :2019-06-30   Max.   :272.0   Max.   :2373711   Max.   :2415841
##     PROD_NBR       PROD_NAME            PROD_QTY         TOT_SALES
##  Min.   :  1.00   Length:246742      Min.   :  1.000   Min.   :  1.700
##  1st Qu.: 26.00   Class :character   1st Qu.:  2.000   1st Qu.:  5.800
##  Median : 53.00   Mode  :character   Median :  2.000   Median :  7.400
##  Mean   : 56.35                      Mean   :  1.908   Mean   :  7.321
##  3rd Qu.: 87.00                      3rd Qu.:  2.000   3rd Qu.:  8.800
##  Max.   :114.00                      Max.   :200.000   Max.   :650.000
```

Table interpretation

```
NO NULL VALUES
If they were present:
 - that means this data could affect averages, totals, or forecasting models.
 - plotting functions may exclude or fail on missing values.
 - reflect data entry issues
DATE
* The transactions span from  2018-07-02 (2nd July, 2018) to 2019-07-01 (1st July, 2019) , so th
is data covers the whole year.
* min, max, and quartiles in summary() helps to see if the dataset covers the entire period cons
istently or if it's skewed to a certain season
* If the data only existed in a part of the year, the analysis would be reflecting seasonality r
ather than overall business trends.

STORE_NBR
* Store numbers range from 1 to 272.
* Since the 1st quartile, median, and 3rd quartile for STORE_NBR fall well within the minimum an
d maximum values, and the distribution appears evenly spread, this suggests that the data is sta
ble and consistent. There are no abnormal spikes or extreme values, indicating that the dataset
accurately reflects the existing store numbers without outliers or invalid entries.


LYLTY_CRD_NBR
* Loyalty card numbers go up to 2.3M
* Suggesting a huge/large customer base

PROD_NBR
* There are 114 distinct product IDs.
* The 1st Qu = 28 means, if you sort all transactions by their PROD_NBR, 25% of those transactio
ns are associated with products that have ID numbers ≤ 28.
* The median product number is 56 , this means half of all transactions involve products with ID
s ≤ 56, and half involve IDs ≥ 56.

* 3rd Quartile (Q3) = 85 means 75% of transactions involve products with IDs ≤ 85, and 25% invol
ve IDs between 85 and 114.
* Max means the remaining 25% involved sold product IDs were between 85 and 114.
* If Max was unexpectedly large, it could indicate extra or invalid products.

PROD_QTY
* Most purchases involve 1 or 2 units (min and max)
* Max = 200 stands out as a potential outlier
* If the max is very far from 3rd Quartile (if the difference is really huge) this could mean a
few reasons: invalid data entry, bulk sale or promotion.
```

Find the transaction that bought 200 packets of chips.

DATE <date>	STORE_… <int>	LYLTY_CARD_… <int>	TXN… <int>	PROD… <int>	PROD_NAME <chr>
2018-08-19	226	226000	226201	4	Dorito Corn Chp Supreme 380g
2019-05-20	226	226000	226210	4	Dorito Corn Chp Supreme 380g

2 rows | 1-7 of 8 columns

```
Observation:
There are two transactions where 200 packets of chips are bought in one transaction
and both of these transactions were by the same customer.

They bought Dorito Corn Chp Supreme 380g, from store number 226 on 2018-08-20 and 2019-05-21.

To identify the customer we can use their LYLTY_CARD_NBR which is 226000.
```

Check if the customer with LYLTY_CARD_NBR 226000 has had other transactions

DATE <date>	STORE_… <int>	LYLTY_CARD_… <int>	TXN… <int>	PROD… <int>	PROD_NAME <chr>
2018-08-19	226	226000	226201	4	Dorito Corn Chp Supreme 380g
2019-05-20	226	226000	226210	4	Dorito Corn Chp Supreme 380g

2 rows | 1-7 of 8 columns

```
Observations:
It appears that customer with LYLTY_CARD_NBR 226000, has only made 2 purchases so far. Where the
y did the bulk purchasing. Since the customer does not appear to be a regular retail customer, w
e could remove them from further analysis.
```

Removing outlier customer basing on the loyalty card number

```
## [1] 0
```

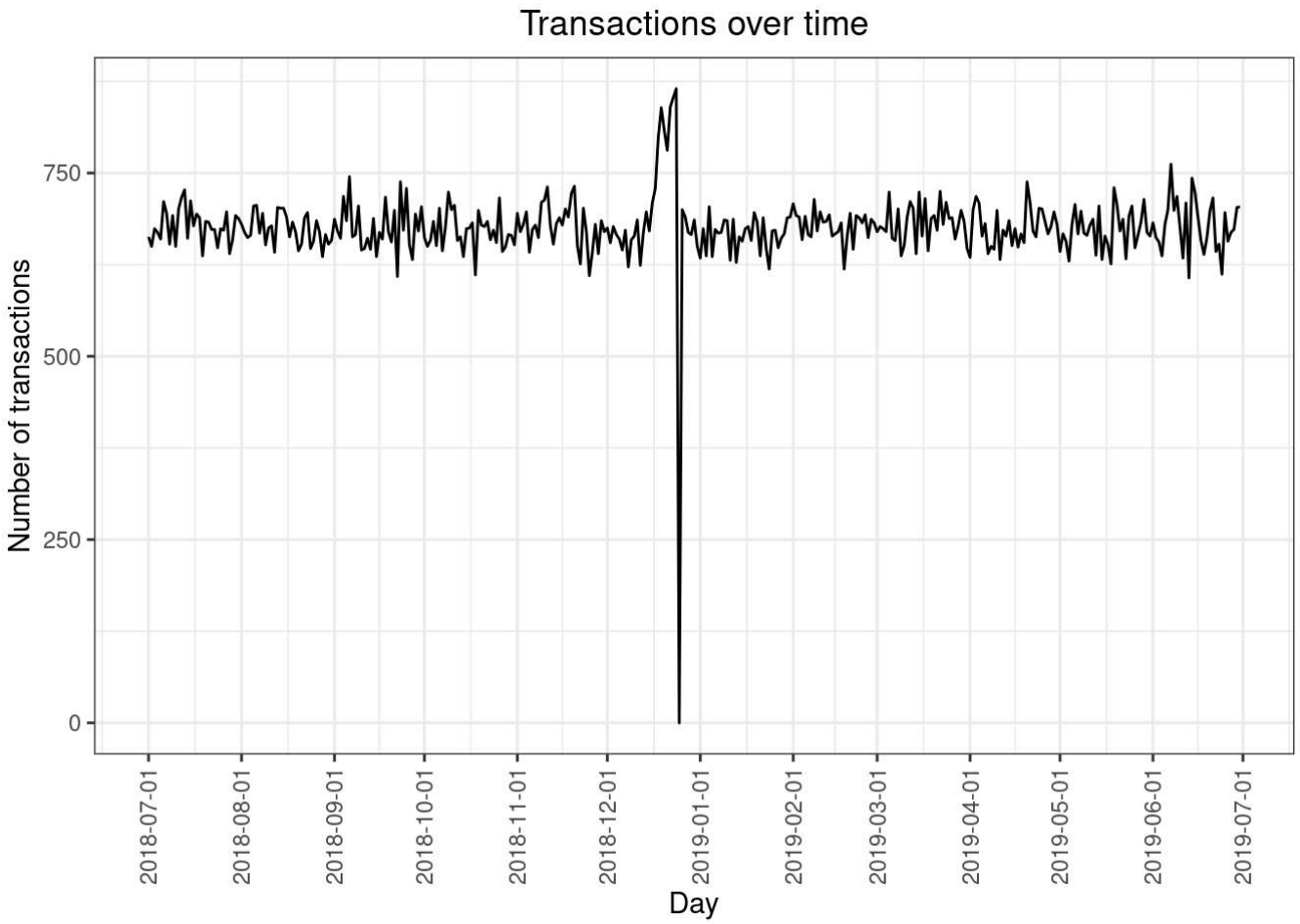
Number of transaction lines over time

	DATE <date>	N <int>
	2018-12-24	865
	2018-12-23	853
	2018-12-22	840
	2018-12-19	839
	2018-12-20	808
	2018-12-18	799
	2018-12-21	781
	2019-06-07	762
	2018-09-06	745
	2019-06-14	743

1-10 of 364 rows Previous **1** 2 3 4 5 6 … 37 Next

```
Observation: Our highest number of transactions is on 2018-12-25 and is 939 . Our lowest number of transactions is on 2018-11-26 and is 648.

- There are 364 rows, so accounting for 364 dates yet we expect 365 dates because the number of days in the period from July 2, 2018, to July 1, 2019, inclusive, is 365 days.

- There is a missing date.
```

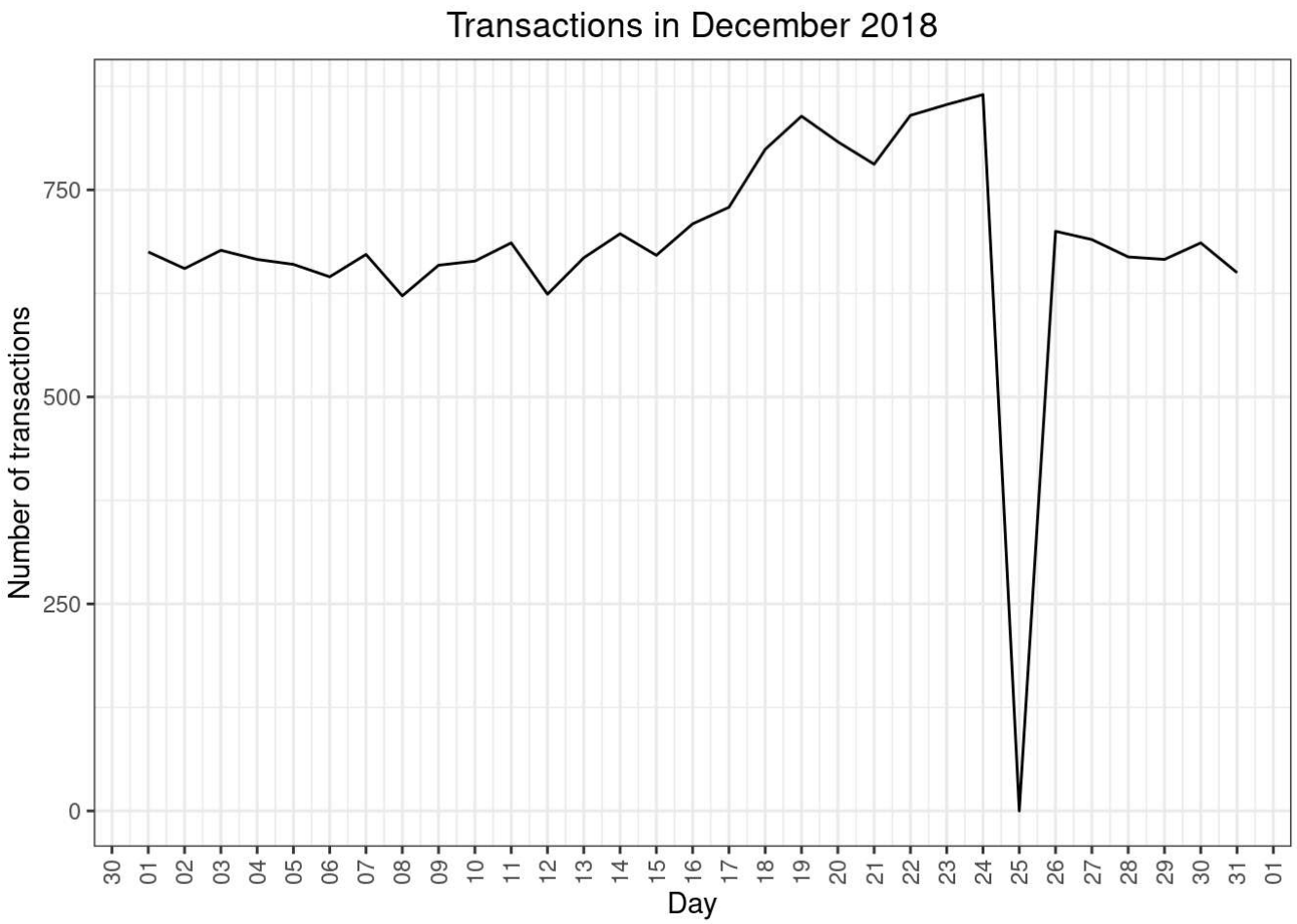
Creating a sequence of dates




```
## [1] "2018-12-25"
```

```
Observation: The date with no transaction is 2018-12-25. Popular holidays celebrated on this dat
e are Christmas day, this could mean that the store was closed hence no transactions.

- A few days before the 2018-12-25, we notice that the sales increase more than they've ever bee
n the previous years.
```

Filtering to December




```
Observations:
The increase in sales occurs in the lead-up to Christmas and there are zero sales on Christmas d
ay itself (12-25-2018). We can now confirm that our data has no outliers.
```

Product Pack Size

```
## [1] "Natural Chip        Compny SeaSalt175g"
## [2] "CCs Nacho Cheese    175g"
## [3] "Smiths Crinkle Cut  Chips Chicken 170g"
## [4] "Smiths Chip Thinly  S/Cream&Onion 175g"
## [5] "Kettle Tortilla ChpsHny&Jlpno Chili 150g"
## [6] "Smiths Crinkle Chips Salt & Vinegar 330g"
```

PACK_SIZE <dbl>	N <int>
380	6416
330	12540
270	6285
250	3169

PACK_SIZE <dbl>	N <int>
220	1564
210	6272
200	4473
190	2995
180	1468
175	66390

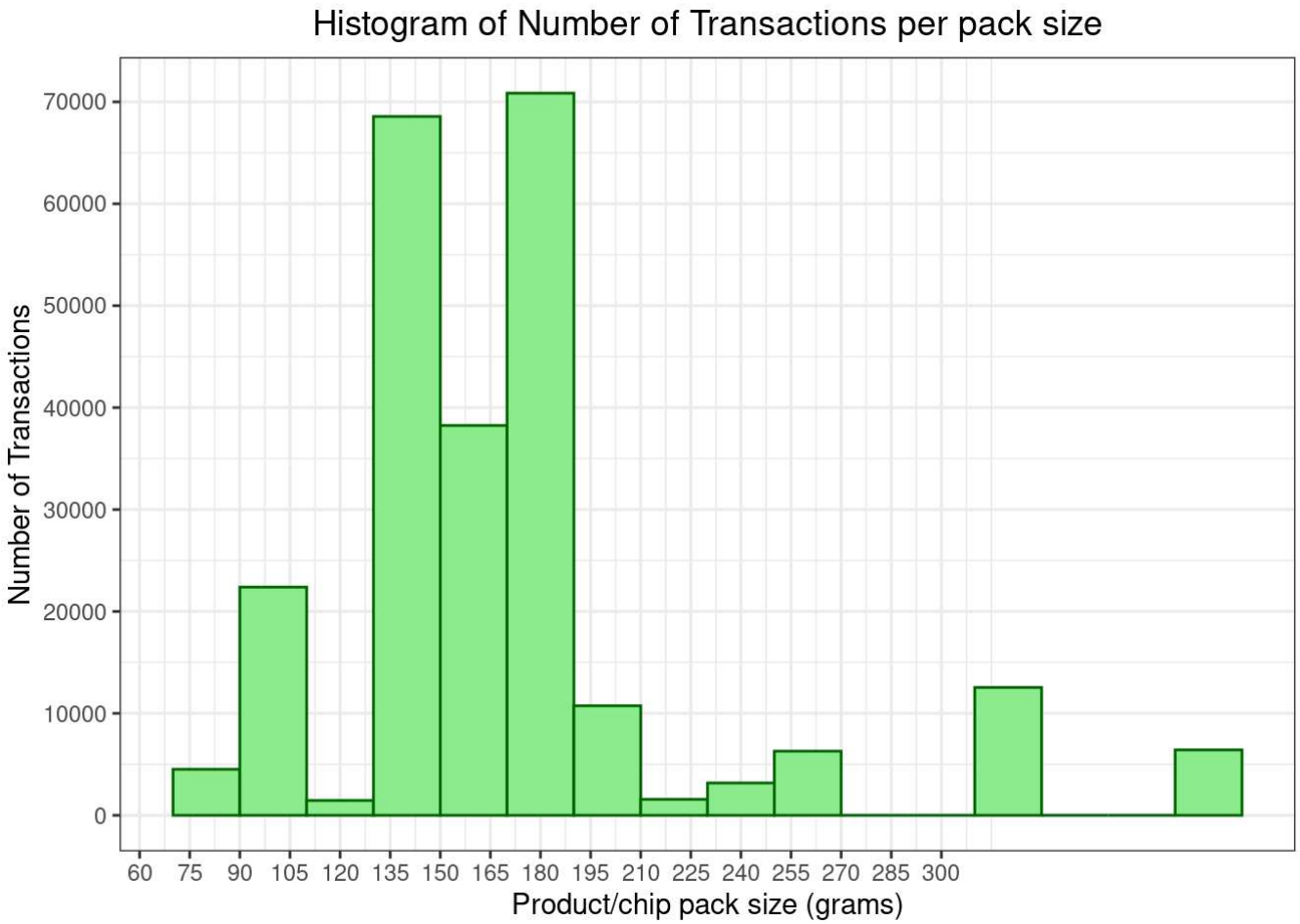
1-10 of 20 rows Previous **1** 2 Next

PACK_SIZE <dbl>	Total_Units <int>
70	2855
90	5692
110	42835
125	2730
134	48019
135	6212
150	76662
160	5604
165	29051
170	38088

1-10 of 20 rows Previous **1** 2 Next

```
Interpretation:
- The largest pack size sold is 380g, and it appears in 6418 rows
- The smallest pack size sold is 70g, and it appears in 1507 rows
-There were 6418 transaction lines/records/rows involving 380g products/chips.
-There were 1507 transaction lines/records/rows involving 70g products/chips.
- Total_Units tells us that we sold 2855 products each weighing 70g.
- Total_Units tells us that we sold 12673 products each weighing 380g.
- Number of transactions by pack size != Number of units sold per pack size.
```

Histogram of PACK_SIZE




```
Interpretation:
- Products with pack sizes ranging between 70 g and 90 g account for approximately 5,000 transac
tion lines.
```

Create Brand Name

```
##  [1] "Natural"    "CCs"        "Smiths"     "Kettle"     "Grain"
##  [6] "Doritos"    "Twisties"   "WW"         "Thins"      "Burger"
## [11] "NCC"        "Cheezels"   "Infzns"     "Red"        "Pringles"
## [16] "Dorito"     "Infuzions"  "Smith"      "GrnWves"    "Tyrrells"
## [21] "Cobs"       "French"     "RRD"        "Tostitos"   "Cheetos"
## [26] "Woolworths" "Snbts"      "Sunbites"
```

PROD_NAME <chr>	BRAND <chr>
Natural Chip Compny SeaSalt175g	Natural
CCs Nacho Cheese 175g	CCs
Smiths Crinkle Cut Chips Chicken 170g	Smiths
Smiths Chip Thinly S/Cream&Onion 175g	Smiths

PROD_NAME <chr>	BRAND <chr>
Kettle Tortilla ChpsHny&Jlpno Chili 150g	Kettle
Smiths Crinkle Chips Salt & Vinegar 330g	Smiths
Grain Waves Sweet Chilli 210g	Grain
Doritos Corn Chip Mexican Jalapeno 150g	Doritos
Grain Waves Sour Cream&Chives 210G	Grain
Smiths Crinkle Chips Salt & Vinegar 330g	Smiths

1-10 of 10,000 rows Previous **1** 2 3 4 5 6 … 1000 Next

Check brand names

BRAND <chr>	N <int>
Kettle	41288
Smiths	30353
Doritos	25224
Pringles	25102
Infuzions	14201
Thins	14075
RRD	11894
Woolworths	11836
Cobs	9693
Tostitos	9471

1-10 of 23 rows Previous **1** 2 3 Next

```
Interpretation:
Now we know that there was 41,288 transaction lines involving chips from the Kettle brand, and i
t had the most. The brand name with the least transaction lines at 1,418 is French.
```

Customer Data Set

	LYLTY_CARD_NBR <int>	LIFESTAGE <chr>	PREMIUM_CUSTOMER <chr>
1	1000	YOUNG SINGLES/COUPLES	Premium

	LYLTY_CARD_NBR <int>	LIFESTAGE <chr>	PREMIUM_CUSTOMER <chr>
2	1002	YOUNG SINGLES/COUPLES	Mainstream
3	1003	YOUNG FAMILIES	Budget
4	1004	OLDER SINGLES/COUPLES	Mainstream
5	1005	MIDAGE SINGLES/COUPLES	Mainstream
6	1007	YOUNG SINGLES/COUPLES	Budget

6 rows

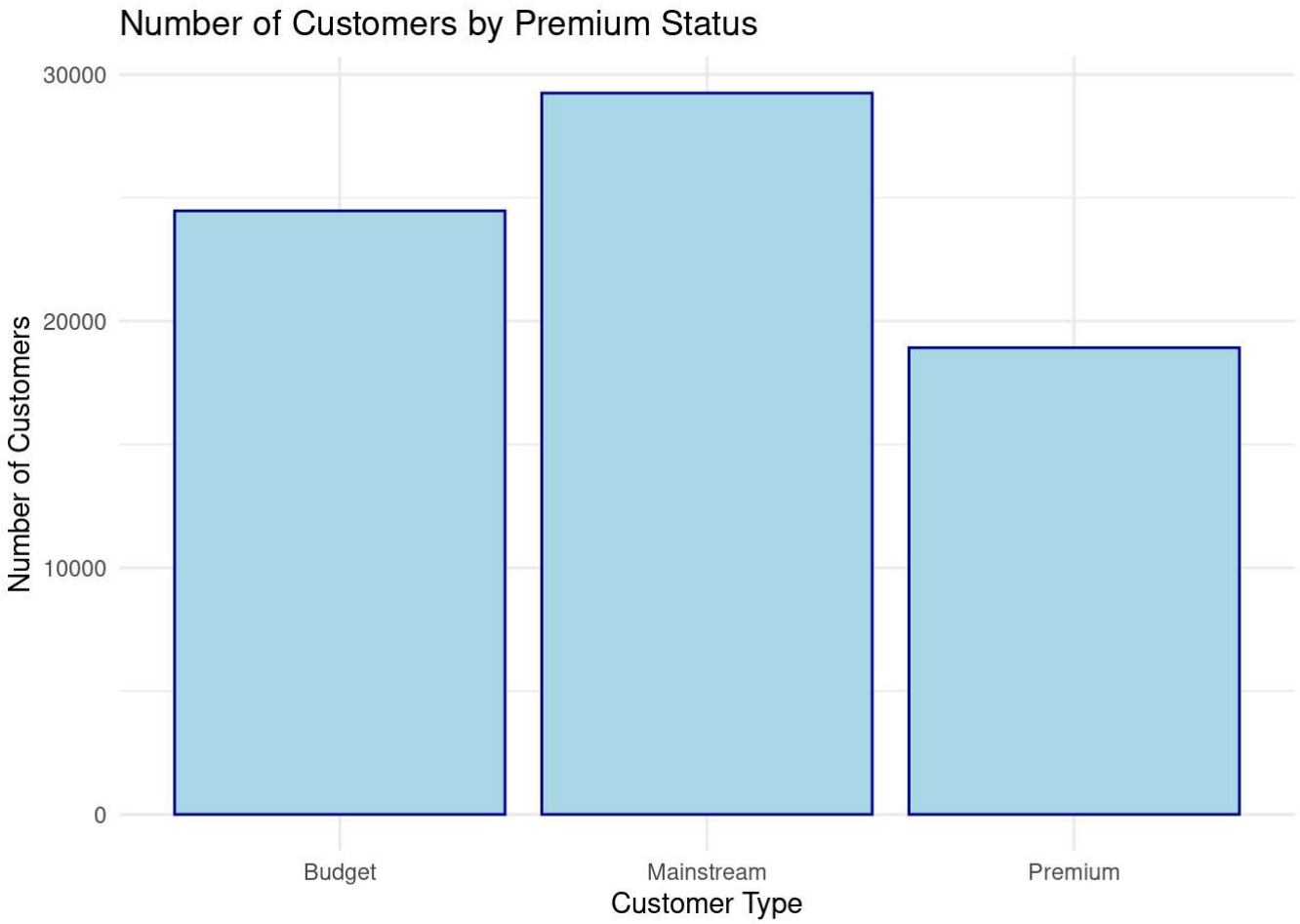
```
## 'data.frame':    72637 obs. of  3 variables:
##  $ LYLTY_CARD_NBR  : int  1000 1002 1003 1004 1005 1007 1009 1010 1011 1012 ...
##  $ LIFESTAGE       : chr  "YOUNG SINGLES/COUPLES" "YOUNG SINGLES/COUPLES" "YOUNG FAMILIES" "O
LDER SINGLES/COUPLES" ...
##  $ PREMIUM_CUSTOMER: chr  "Premium" "Mainstream" "Budget" "Mainstream" ...
```

```
## [1] "YOUNG SINGLES/COUPLES"  "YOUNG FAMILIES"         "OLDER SINGLES/COUPLES"
## [4] "MIDAGE SINGLES/COUPLES" "NEW FAMILIES"           "OLDER FAMILIES"
## [7] "RETIREES"
```

```
## [1] "Premium"    "Mainstream" "Budget"
```

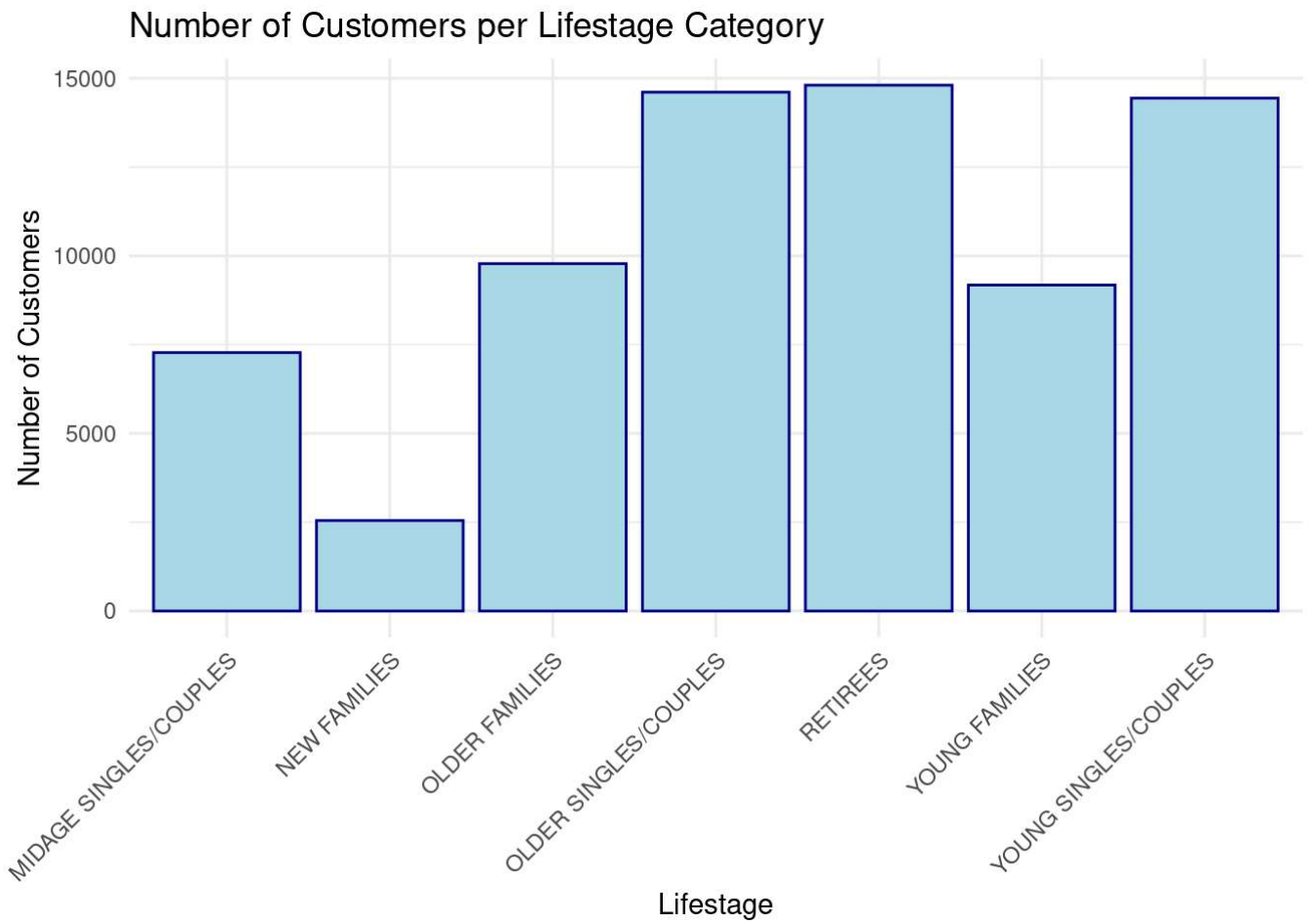
```
Interpretation:
-Categories of LIFESTAGE: YOUNG SINGLES/COUPLES, YOUNG FAMILIES, OLDER SINGLES/COUPLES, MIDAGE S
INGLES/COUPLES, NEW FAMILIES, OLDER FAMILIES, RETIREES
- Categories of PREMIUM_CUSTOMER: Premium, Mainstream, Budget
```

Bar Chart of Premium Customer Types




```
The Mainstream category has the most number of customers. Followed by Budget and lastly, Premiu
m.
```

Bar Chart of Lifestage Type



Observation: Majority of the customer base are Retirees and OLDER SINGLES/COUPLES. New Families represent the smallest customer segment.

Merge Transaction Data to Behaviour(Customer) Data

LYLTY_CARD_… <int>	DATE <date>	STORE_… <int>	TXN… <int>	PROD… <int>	PROD_NAME <chr>
1000	2018-10-17	1	1	5	Natural Chip Compny SeaSalt175g
1002	2018-09-16	1	2	58	Red Rock Deli Chikn&Garlic Aioli 150g
1003	2019-03-07	1	3	52	Grain Waves Sour Cream&Chives 210G
1003	2019-03-08	1	4	106	Natural ChipCo Hony Soy Chckn175g
1004	2018-11-02	1	5	96	WW Original Stacked Chips 160g
1005	2018-12-28	1	6	86	Cheetos Puffs 165g

6 rows | 1-6 of 12 columns

Guiding notes

```
* we used a left join (all.x = TRUE), which keeps all rows from transactionData and only adds ma
tching details from the other table.
* finds a shared column: LYLTY_CARD_NBR
* looks at each LYLTY_CARD_NBR in transactionData.
* finds matching rows in behaviourData with the same LYLTY_CARD_NBR.
* copies over the extra columns (like LIFESTAGE or PREMIUM_CUSTOMER) from behaviourData into dat
a.
```

Check for missing customer details

```
##   LYLTY_CARD_NBR             DATE        STORE_NBR           TXN_ID
##                0                0                0                0
##         PROD_NBR        PROD_NAME         PROD_QTY        TOT_SALES
##                0                0                0                0
##        PACK_SIZE            BRAND        LIFESTAGE PREMIUM_CUSTOMER
##                0                0                0                0
```

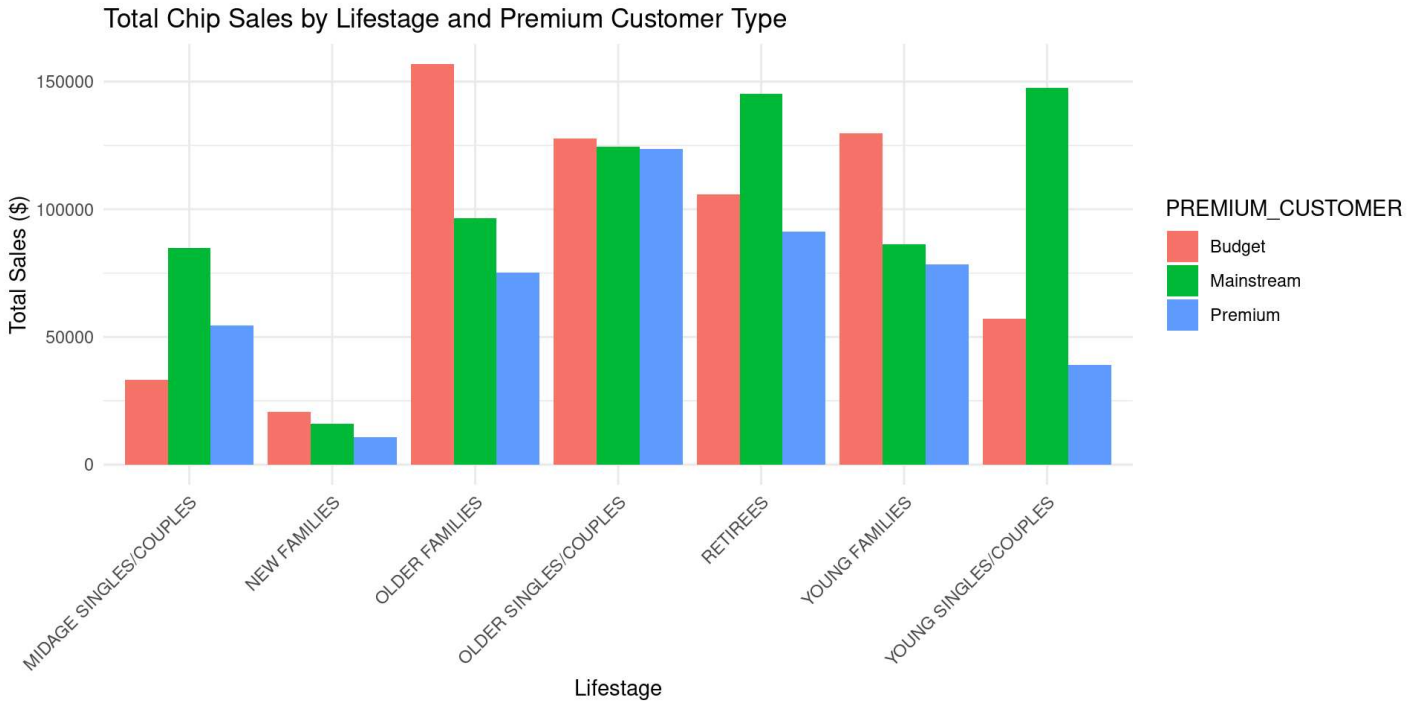
```
Interpretation: There are 0 missing values in every column.
This can signal that our dataset is clean.
```

Saving “data” dataset for next steps

```
## [1] "DATA EXPLORATION IS NOW COMPLETE"
```

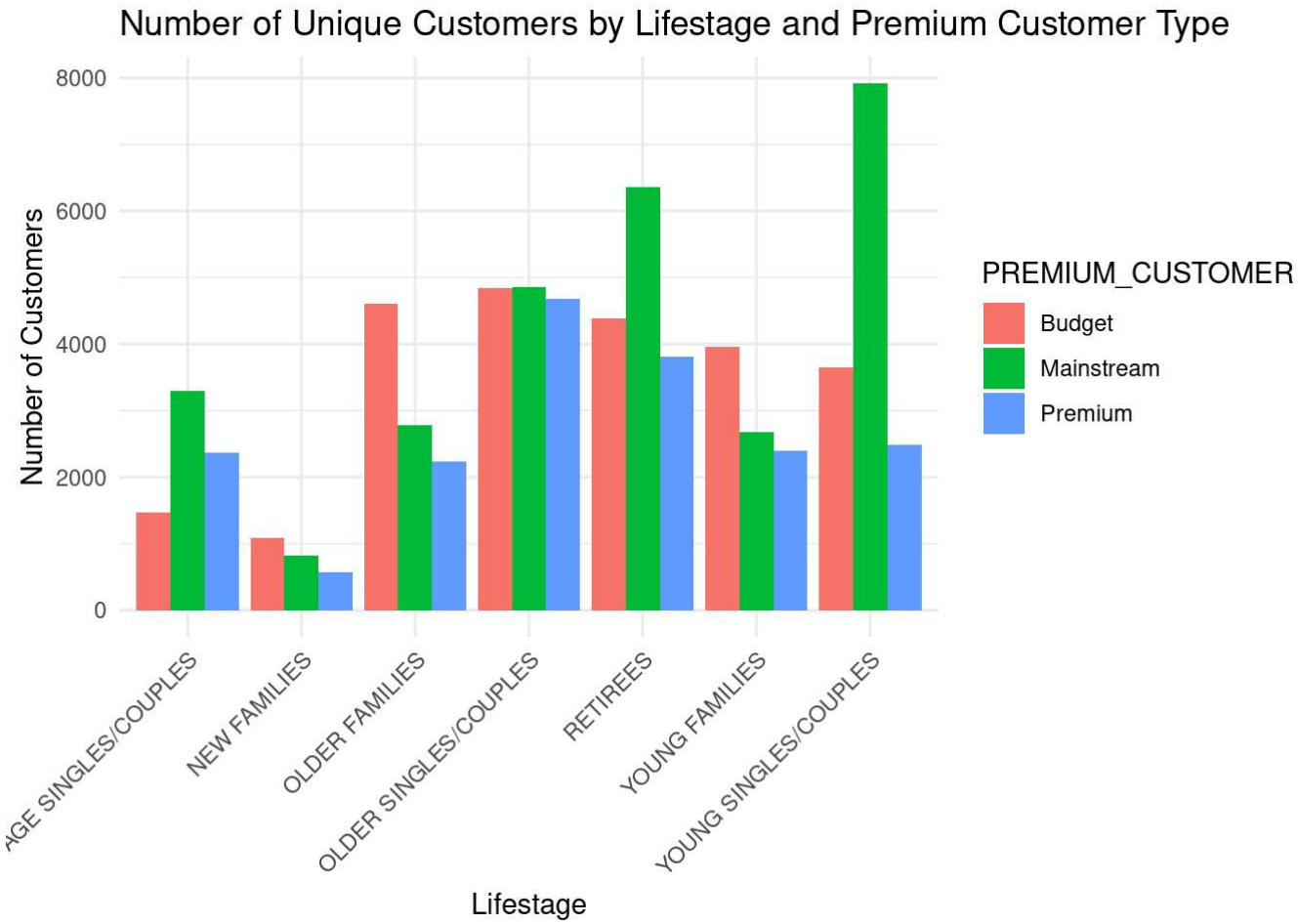
Data Analytics on Customer Segments

```
sales_by_segment <- data[, .(Total_Sales = sum(TOT_SALES)), by = .(LIFESTAGE, PREMIUM_CUSTOMER)]

ggplot(sales_by_segment, aes(x = LIFESTAGE, y = Total_Sales, fill = PREMIUM_CUSTOMER)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(
    title = "Total Chip Sales by Lifestage and Premium Customer Type",
    x = "Lifestage",
    y = "Total Sales ($)"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
Interpretation:
- Mainstream and Budget customers are the primary audience driving chip sales.
- Older Families (Budget) have the highest total sales out of all combinations, peaking above $1
50,000. Hence they have a strong purchasing power.
- Retirees and Young Singles/Couples also show high sales, particularly for Mainstream customer
s.
- All segments within New Families have low total sales.
```

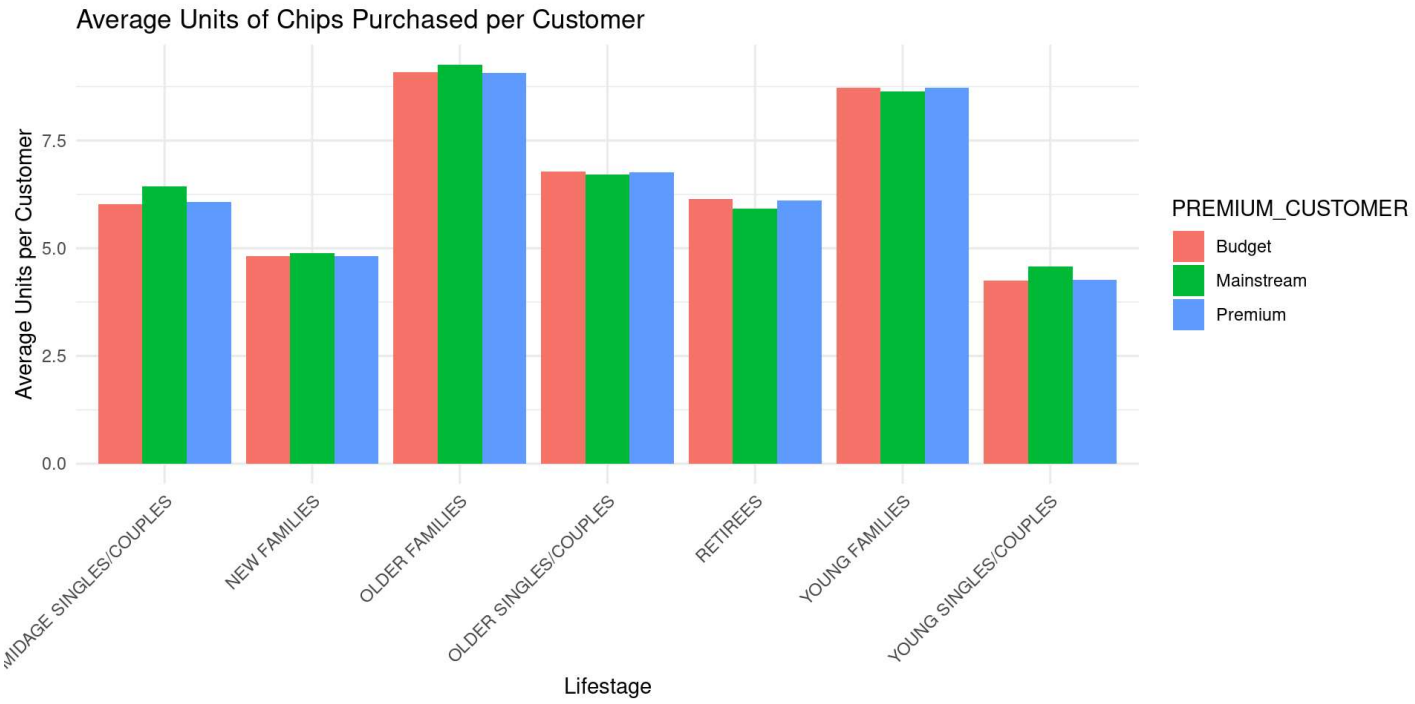
Unique customers per segment




```
Observations
- Mainstream Young Singles/Couples make up the largest share of our customer base. That also exp
lains why this segment also had one of the highest total sales in the previous chart.
- Mainstream retirees have high customer counts too.
- Segments with higher sales also tend to have larger customer counts, indicating that total sal
es could be driven more by customer volume than by higher individual spending. Let us evaluate t
his in the next step.
```

Total units purchased and unique customers per segment

```
units_per_segment <- data[, .(
  Total_Units = sum(PROD_QTY),
  Num_Customers = uniqueN(LYLTY_CARD_NBR)
), by = .(LIFESTAGE, PREMIUM_CUSTOMER)]

units_per_segment[, Avg_Units_Per_Customer := Total_Units / Num_Customers]

ggplot(units_per_segment, aes(x = LIFESTAGE, y = Avg_Units_Per_Customer, fill = PREMIUM_CUSTOME
R)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(
    title = "Average Units of Chips Purchased per Customer",
    x = "Lifestage",
    y = "Average Units per Customer"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Observation:
- Older families and young families in general buy more chips per customer.

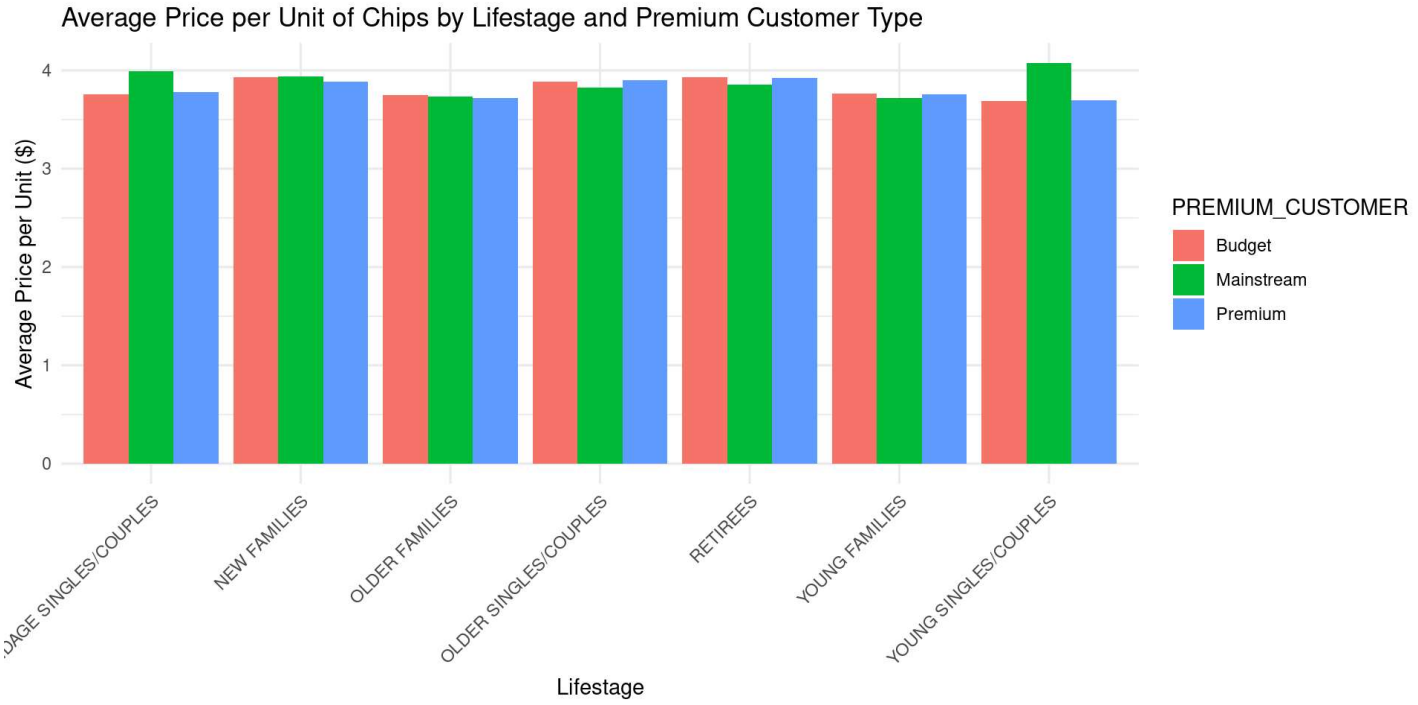
- The above chart was done to check if in this case, the higher sales are driven by more units of chips being bought per customer.

- Customers in these groups are more likely to buy in bulk or make frequent purchases, leading to higher sales per customer.Older Families customers purchase on average 8-9 units of chips per person/customer.

- While the previous chart focused on the average number of units purchased per customer, analyzing the price per unit will help determine if higher sales are also influenced by customers buying more expensive products, not just larger quantities.

analyzing average price per unit sales

```
price_per_segment <- data[, .(
  Total_Sales = sum(TOT_SALES),
  Total_Units = sum(PROD_QTY)
), by = .(LIFESTAGE, PREMIUM_CUSTOMER)]

price_per_segment[, Avg_Price_Per_Unit := Total_Sales / Total_Units]

ggplot(price_per_segment, aes(x = LIFESTAGE, y = Avg_Price_Per_Unit, fill = PREMIUM_CUSTOMER)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(
    title = "Average Price per Unit of Chips by Lifestage and Premium Customer Type",
    x = "Lifestage",
    y = "Average Price per Unit ($)"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



```
Observation:
* The average price of chips purchased across all premium customer types and lifestage segments
is roughly the same range. Except for Midage Singles/Couples and Young singles/couples.

* This indicates that product pricing is not the primary driver of high total sales. Instead, th
e strong sales performance is more likely due to customers purchasing larger quantities per pers
on (around 8–9 units on average), rather than buying more expensive chip products. Like we saw i
n the previous chart.

* It would be smart to validate this further, confirming that sales growth is volume-driven rath
er than price-driven.

* Since Mainstream midage and young singles and couples pay more on average price per unit of ch
ips compared to their budget and premium counterparts. This may be due
to premium shoppers being more likely to buy healthy snacks and when they buy
chips, this is mainly for entertainment purposes rather than their own consumption.
This is also supported by there being fewer premium midage and young singles and
couples buying chips compared to their mainstream counterparts.

* Even though we said that the average price of chips is roughly the same range, we can proceed
to perform a t-test. (Check the difference between the different average/mean prices of each cus
tomer segment)
```

Statistical Analysis

```
##
##  Welch Two Sample t-test
##
## data:  TOT_SALES by PREMIUM_CUSTOMER
## t = 24.24, df = 24455, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Mainstream and group Premium i
s not equal to 0
## 95 percent confidence interval:
##  0.5866125 0.6898246
## sample estimates:
## mean in group Mainstream    mean in group Premium
##                 7.582377                 6.944158
```

```
##
##  Welch Two Sample t-test
##
## data:  TOT_SALES by PREMIUM_CUSTOMER
## t = -28.678, df = 23855, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Budget and group Mainstream is
not equal to 0
## 95 percent confidence interval:
##  -0.8138943 -0.7097560
## sample estimates:
##     mean in group Budget mean in group Mainstream
##                 6.820552                 7.582377
```

Interpreting results

```
* Mainstream vs Premium customer's p-value = 2.2e-16 (below 0.5)
   - Mean Mainstream: 7.59
   - Mean Premium: 6.90
   - Difference: 0.69
   - Mainstream customers pay 0.69$ more than Premium customers on average.
   - since p-value is less than 0.5, this means the observed results are statistically significa
nt and it is unlikely they occurred by random chance alone. E.g mainstream customers might choos
e more expensive products, more items per transaction increasing the average price, this value s
inifies that there are reasons causing this.

* Budget vs Mainstream customer's p-value = 2.2e-16 (below 0.5)
   - Mean Budget: 6.78
   - Mean Mainstream: 7.59
   - Difference: 0.81
   - Mainstream customers pay 0.81$ more than Budget customers on average.
   - since p-value is less than 0.5, this also means the observed results are statistically sign
ificant and it is unlikely they occurred by random chance alone. There are valid market factors
that could be influencing this.
   - Mainstream customers have the highest average spend per transaction, reinforcing why they c
ontribute so strongly to total sales.

* Next steps: let's find out if mainstreams tend to buy a particular brand of chips.
```

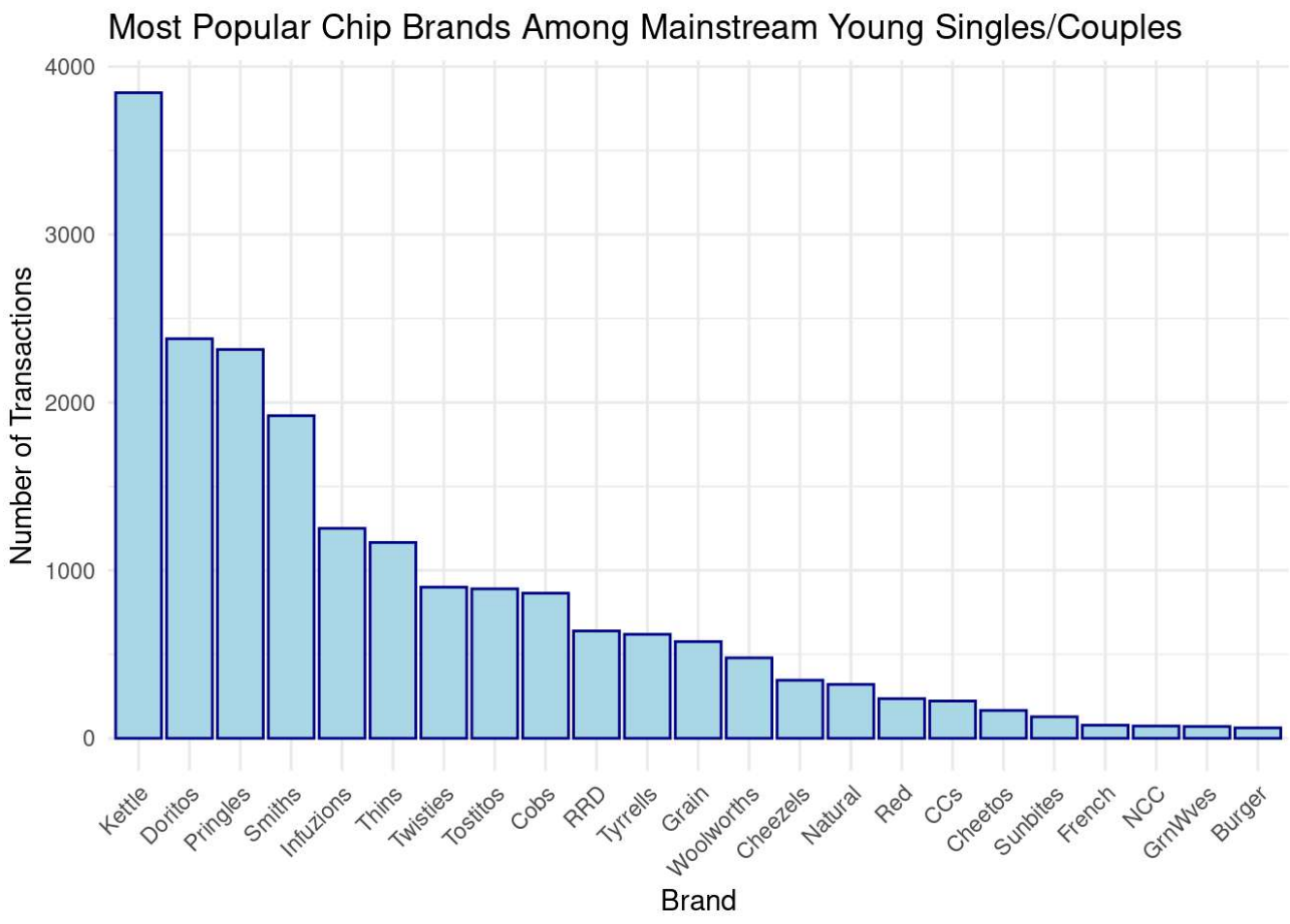
Mainstream Customers Analysis

```
mainstream_young <- data[
  PREMIUM_CUSTOMER == "Mainstream" &
  LIFESTAGE == "YOUNG SINGLES/COUPLES"
]

brand_pref <- mainstream_young[, .N, by = BRAND][order(-N)]
head(brand_pref)
```

BRAND <chr>	N <int>
Kettle	3844
Doritos	2379
Pringles	2315
Smiths	1921
Infuzions	1250
Thins	1166

6 rows

```
ggplot(brand_pref, aes(x = reorder(BRAND, -N), y = N)) +
  geom_bar(stat = "identity", fill = "lightblue", color = "darkblue") +
  labs(
    title = "Most Popular Chip Brands Among Mainstream Young Singles/Couples",
    x = "Brand",
    y = "Number of Transactions"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



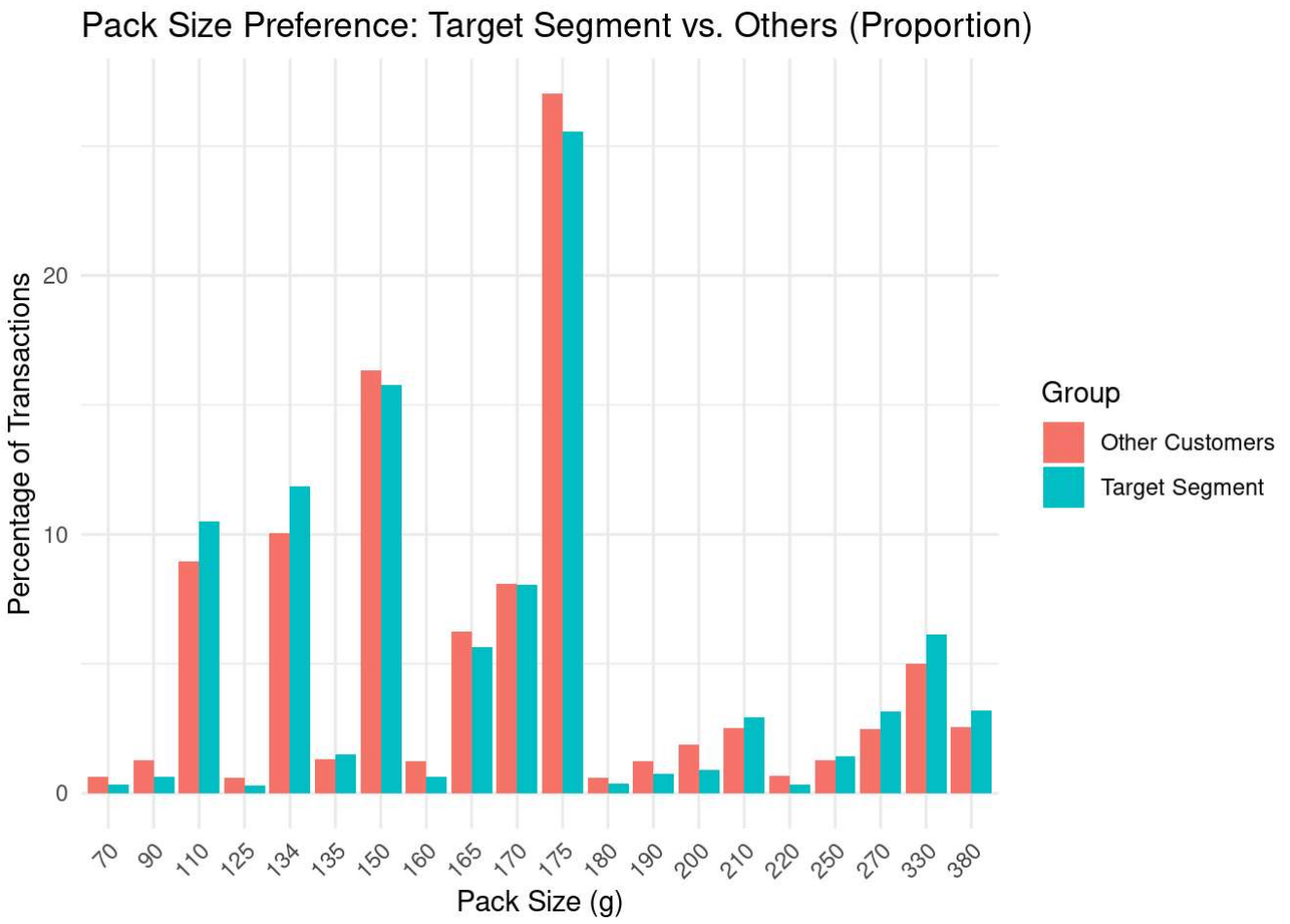
```
#Affinity Analysis

# Share of each brand in this segment
brand_pref[, Segment_Share := N / sum(N)]

# Overall share of each brand
overall_brand <- data[, .N, by = BRAND]
overall_brand[, Overall_Share := N / sum(N)]

# Merge both
brand_affinity <- merge(brand_pref, overall_brand[, .(BRAND, Overall_Share)], by = "BRAND")

# Calculate Lift
brand_affinity[, Lift := Segment_Share / Overall_Share]

# Sort by Lift
brand_affinity[order(-Lift)]
```

BRAND <chr>	N <int>	Segment_Share <dbl>	Overall_Share <dbl>	Lift <dbl>
Tyrrells	619	0.031672124	0.026108454	1.2130984
Twisties	900	0.046049939	0.038315636	1.2018576
Doritos	2379	0.121725338	0.102229067	1.1907116
Tostitos	890	0.045538273	0.038384534	1.1863703
Kettle	3844	0.196684404	0.167334036	1.1753999
Pringles	2315	0.118450675	0.101734619	1.1643104
Grain	576	0.029471961	0.025419470	1.1594247
Cobs	864	0.044207941	0.039284267	1.1253345
Infuzions	1250	0.063958248	0.057554511	1.1112639
Thins	1166	0.059660254	0.057043852	1.0458665

1-10 of 23 rows Previous **1** 2 3 Next

```
Observation: Mainstream Young Singles/Couples most frequently buy Kettle.
- If lift > 1: means the segment is more likely than average to buy that brand.
- If lift < 1: means they’re less likely than average.
- From the table, they are most likely to purchase Tyrrells, Twisties, Tostitos and Kettle.
- Note: The Lift values show how much more likely this segment is to buy a brand compared to the
overall population.
```

Target Segment Analysis




```
## Average Pack Size - Target Segment: 178.3442
```

```
## Average Pack Size - Other Customers: 175.346
```

```
Observation:
- Our target segment : Mainstream customers (who are also young singles and couples) show a clea
r spike at 175 g, making it the most commonly purchased pack overall.
- Smaller packs (70–110 g) are less popular among the target segment compared to others.
- The target segment buys larger packs (270–330 g) at a higher proportion than other customers,
suggesting a preference for bulk or shareable sizes.
- findings reinforce that high sales for this segment are driven by larger pack purchase not hig
her unit prices like we thought earlier.
```

-