Quantium data analysis: chips in the region

Import data

Created table with data types and imported data

```
CREATE TABLE qvi_transaction_data
(
dates DATE,
store_num INT,
loyalty_card_num INT,
txin_id INT,
prod_num INT,
product_name VARCHAR(255),
product_qty INT,
total_sales FLOAT
);
```

Display the table to check if table is created

```
select * from qvi_transaction_data;
```

Imported csv file using load data command

```
LOAD DATA INFILE 'C:/ProgramData/MySQL/MySQL Server
8.0/Uploads/QVI_transaction_data.csv'
INTO TABLE qvi_transaction_data
FIELDS TERMINATED BY ','
ENCLOSED BY '"'
LINES TERMINATED BY '\r\n'
IGNORE 1 LINES;
```

Checked data type of each column after importing

```
SHOW FIELDS FROM qvi_transaction_data;
```

Displayed top 10 rows

```
SELECT DISTINCT * FROM qvi_transaction_data LIMIT 10;
```

Data processing

Changing date from integer to proper date

Change in excel and import - select column -> format cells -> date
Use cast/from_unixtime in mysql to convert integer to datetime

Extract size and put it in a new column named pack size from the product name
alter table qvi_transaction_data add packsize varchar(25);
update qvi_transaction_data set packsize = REGEXP_REPLACE(product_name, '[^0-9]', '');
update qvi_transaction_data set packsize = concat(packsize, 'g');

Clean product names remove sizes, spaces and special characters
Removed size of a particular brand that had the pack size in centre rather than end
update qvi_transaction_data set product_name = concat(left(product_name,7),
right(product_name,17)) where prod_num = 63;

Removed pack sizes from product name
update qvi_transaction_data set product_name = left(product_name,length(product_name)-4)
where prod_num NOT LIKE 63;

Remove extra white space between words
UPDATE qvi_transaction_data set product_name = REGEXP_REPLACE(product_name,
'[[:space:]]+', ' ');

Remove special characters from product name
update qvi_transaction_data set product_name = REPLACE(product_name, '&', '');
update qvi_transaction_data set product_name = REPLACE(product_name, '/', '');

Most occurred words from product name
SELECT product_name,COUNT(*) FROM qvi_transaction_data GROUP BY product_name
ORDER BY COUNT(*) DESC;

Remove salsa from product name so there are only chips
update qvi_transaction_data set product_name = replace(product_name,'salsa','');

check for nulls and possible outliers
select total_sales from qvi_transaction_data where total_sales IS NULL;
NO NULL VALUES

Frequency of values in product quantity
SELECT product_qty, COUNT(*) AS freq FROM qvi_transaction_data GROUP BY product_qty;
Outliers found with 200 product quantity
select * from qvi_transaction_data where product_qty = 200;

Remove outliers
delete from qvi_transaction_data where loyalty_card_num = 226000;

Find missing date from the date range
Select and make duplicate of date column
Write formula to subtract one row from previous row in date column

Find row which gives value as 0
Missing date is 25th dec since it is christmas

Create Column brandname which contains the brand of the product, by extracting it from the product name

```
alter table qvi_transaction_data add brand_name varchar(25);
update qvi_transaction_data set brand_name = substring_index(product_name,' ',1);
```

Change brand name RRD into red, dorito to doritos, infzns to infuzions, Smith to Smiths, GrnWavs to Grain

```
update qvi_transaction_data set brand_name = REPLACE(brand_name,'RRD','red');
update qvi_transaction_data set brand_name = REPLACE(brand_name,'Dorito','Doritos');
update qvi_transaction_data set brand_name = REPLACE(brand_name,'Infzns','Infuzions');
```

Import purchase behaviour table

```
create table qvi_purchase_behaviour (LYLTY_CARD_NBR mediumint, LIFESTAGE
VARCHAR(25), PREMIUM_CUSTOMER VARCHAR(10));
LOAD DATA INFILE 'C:/ProgramData/MySQL/MySQL Server
8.0/Uploads/QVI_purchase_behaviour.csv' INTO TABLE qvi_transaction_data
FIELDS TERMINATED BY ','
ENCLOSED BY '"'
LINES TERMINATED BY '\r\n'
IGNORE 1 LINES;
```

Merge transaction data and purchase behaviour table to create a new final table customerdata

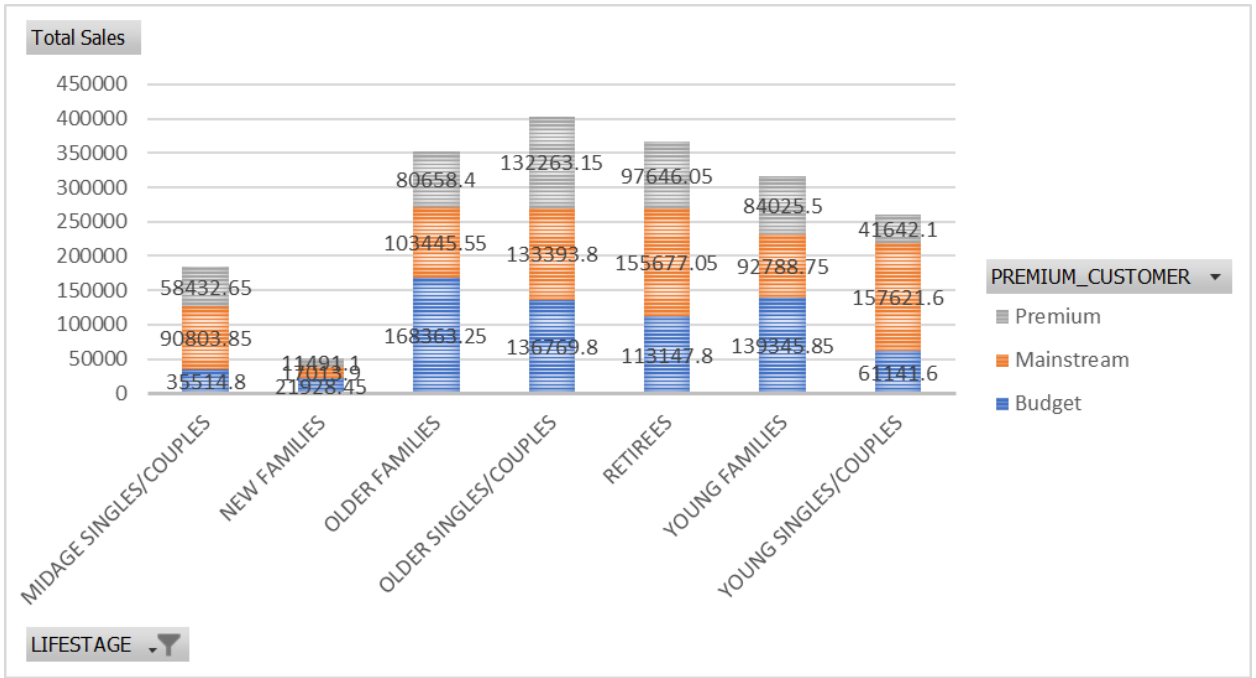
```
create table customerdata as select * from qvi_transaction_data join qvi_purchase_behaviour
on qvi_transaction_data.loyalty_card_num = qvi_purchase_behaviour.LYLTY_CARD_NBR;
alter table customerdata drop LYLTY_CARD_NBR;
```

Data Analysis

Who spends the most on chips (total sales), describing customers by lifestage and how premium their general purchasing behaviour is

Based on lifestyle the segment spending the most on chips is older singles/couples and the segment spending the least is new families. When it comes to being a premium customer there is a similar pattern with older singles/couples being the segment with highest premium customer and new families segment being the lowest

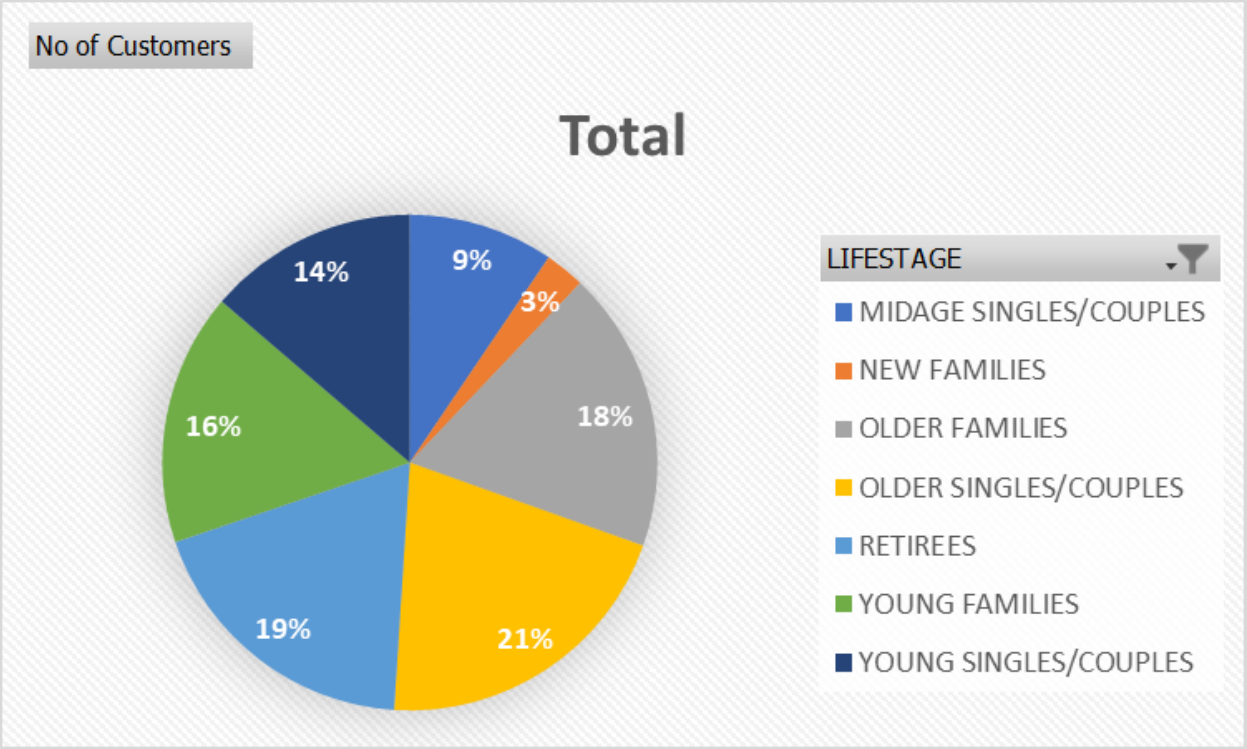
Total Sales	Premium status			
Lifestage	**Budget**	**Mainstream**	**Premium**	**Grand Total**
MIDAGE SINGLES/COUPLES	35514.8	90803.85	58432.65	184751.3
NEW FAMILIES	21928.45	17013.9	11491.1	50433.45
OLDER FAMILIES	168363.25	103445.55	80658.4	352467.2
OLDER SINGLES/COUPLES	136769.8	133393.8	132263.15	402426.75
RETIREES	113147.8	155677.05	97646.05	366470.9
YOUNG FAMILIES	139345.85	92788.75	84025.5	316160.1
YOUNG SINGLES/COUPLES	61141.6	157621.6	41642.1	260405.3
Grand Total	**676211.55**	**750744.5**	**506158.95**	**1933115**



How many customers are in each segment

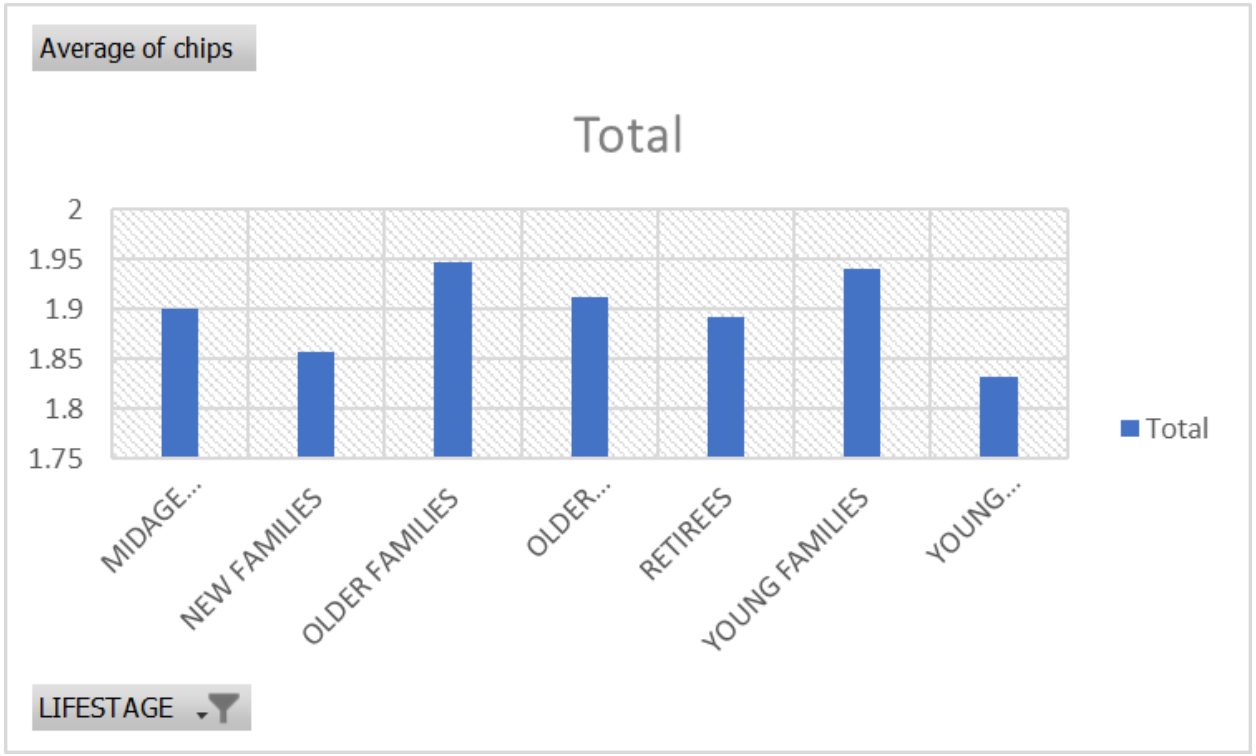
Number of Customers based on segments are represented below

Lifestage	No of Customers
MIDAGE SINGLES/COUPLES	25110
NEW FAMILIES	6919
OLDER FAMILIES	48594
OLDER SINGLES/COUPLES	54479
RETIREES	49763
YOUNG FAMILIES	43592
YOUNG SINGLES/COUPLES	36377
Grand Total	**264834**



How many chips are bought per customer by segment
Most chips bought per customer belongs to older families segment

Lifestage	Average of chips
MIDAGE SINGLES/COUPLES	1.900477897
NEW FAMILIES	1.85677121
OLDER FAMILIES	1.946577767
OLDER SINGLES/COUPLES	1.912718662
RETIREES	1.892289452
YOUNG FAMILIES	1.939828409
YOUNG SINGLES/COUPLES	1.831761828
Grand Total	**1.905812698**



What's the average chip price by customer segment

Average of total_sales	Premium status			
Lifestage	**Budget**	**Mainstream**	**Premium**	**Grand Total**
MIDAGE SINGLES/COUPLES	7.074661355	7.647283982	7.112055745	7.357678216
NEW FAMILIES	7.297321131	7.317806452	7.231655129	7.289124151
OLDER FAMILIES	7.26957038	7.262394693	7.208078642	7.253306993
OLDER SINGLES/COUPLES	7.430314554	7.282115952	7.44976625	7.386823363
RETIREES	7.44344451	7.252261716	7.456173641	7.3643249
YOUNG FAMILIES	7.287200607	7.189025335	7.266756032	7.252709213
YOUNG SINGLES/COUPLES	6.615624324	7.558338928	6.629851934	7.158514996
Grand Total	**7.258837768**	**7.361106209**	**7.263111108**	**7.299346005**