Sales data analysis

Please note that the analysis below was ran on a simulated dataset and should be only considered in the context of the portfolio.

Scenario: The business needs actionable insight based on some sales data.

Executive Summary

The goal of the present analysis was to reach some actionable insights for the business based on the product sales data. Based on the commonalities across measures 3 action groups were created such as marketing, growth and customer insights.

- **Marketing actions**: The analysis revealed that medium shelf location outperforms both good and bad shelf locations in terms of product sales. This means that medium shelf location could be beneficial to prioritise in the future. Furthermore, the data suggests that as the product price increased so did product sales. However, this does not necessarily mean that increasing the price of the product would lead to even higher sales. It is potentially reflective of a contributing factor that wasn't measured. For instance, it is possible that there is less and less selection/competition for our products the more high-quality product it becomes (which would also increase their price) meaning that most people will end up buying our product regardless of the price. Nonetheless, I would suggest no need to decrease the price point of the product either at the moment, but more data on selection would help with gaining further insights. Finally, in our sample competitor product price and advertising costs did not seem to have an effect on the product sales, meaning that new advertising strategies should be considered for the future.
- **Growth actions**: The analysis suggests that the product performs better in non-US and non-urban areas regardless of the population size. According to these results the expansion to less urban areas and outside US could be beneficial for the business in the long-run. These growth oriented actions could include opening of new sites, stocking the products in stores or free/decreased shipping to less urban areas and outside US certain areas.
- **Customer insights**: The results suggest regardless of the average income of the customers increases in education level and age are associated with higher product sales volume. The average age of the customer group is 50, ranging between 25 to 80. It would be beneficial to target these customers from this age group (specifically from the higher ends) and customers with higher education level.

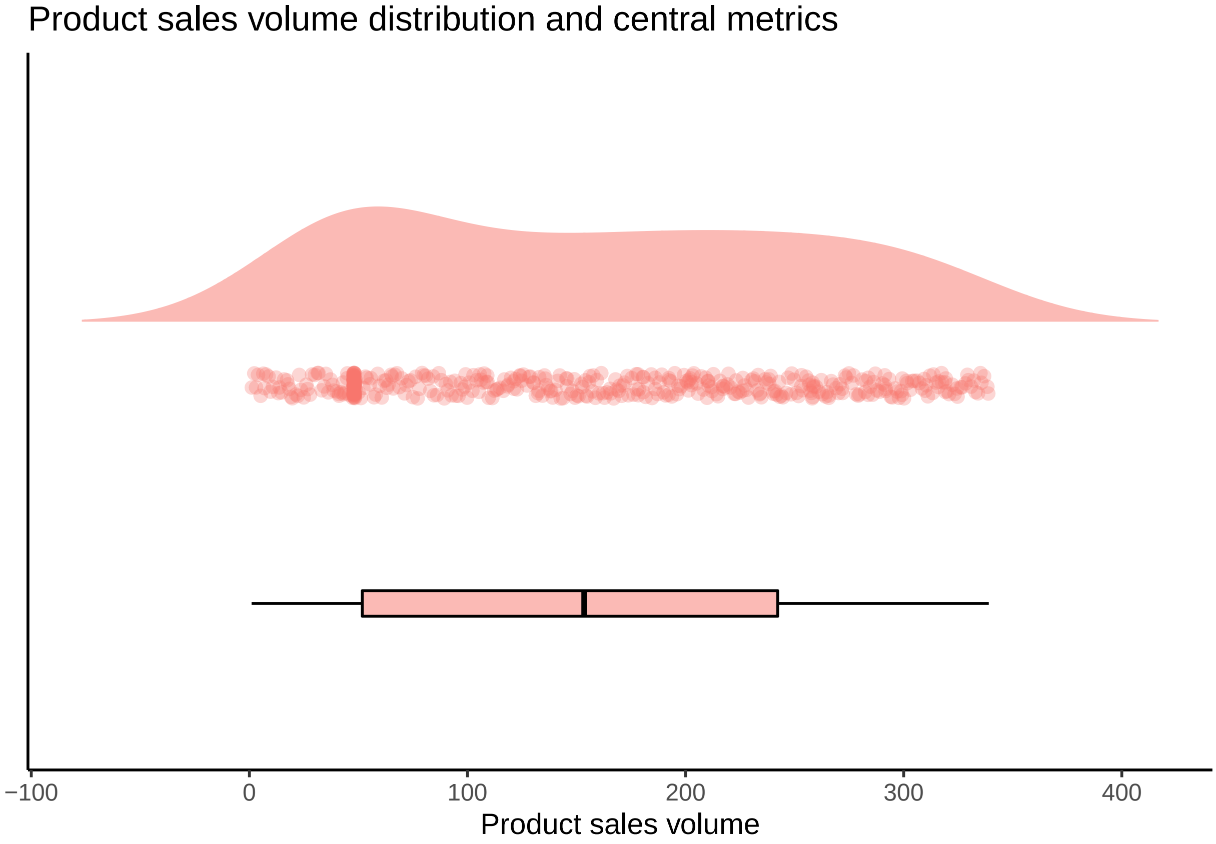
Please note that predictions about the impact is non-causal meaning that experimentation or causal inference would be needed to confirm the above hypothesis and reliably assume any kinds of causality.

Analysis

One observation had NaN on the categorical independent variables (shelf_location, is_US and is_urban) so it was filtered out from the data set, leaving the final data set with N = 464.

Summary of main descriptive statistics across variables:

```
##  product_sales_volume competitor_product_price average_income
##  Min.   :  1.00       Min.   :-60.0            Min.   : 21.00
##  1st Qu.: 51.75       1st Qu.:116.0            1st Qu.: 46.75
##  Median :153.50       Median :125.0            Median : 74.50
##  Mean   :157.44       Mean   :124.3            Mean   : 71.28
##  3rd Qu.:242.25       3rd Qu.:133.2           3rd Qu.: 89.00
##  Max.   :339.00       Max.   :175.0           Max.   :120.00
##  advertising_costs   population    product_price   shelf_location
##  Min.   : 0.00     Min.   : 10.0   Min.   : 24.0   Bad   :156
##  1st Qu.: 0.00     1st Qu.:152.0   1st Qu.: 91.0   Good  : 87
##  Median : 8.00     Median :303.5   Median :111.5   Medium:221
##  Mean   : 7.11     Mean   :280.4   Mean   :112.2   Nan   :  0
##  3rd Qu.:11.00     3rd Qu.:380.0   3rd Qu.:129.0
##  Max.   :29.00     Max.   :509.0   Max.   :191.0
##  representative_age education_level is_urban  is_US
##  Min.   :25.00      Min.   :10.00   Nan:  0   Nan:  0
##  1st Qu.:33.00      1st Qu.:11.00   No :118   No :142
##  Median :50.00      Median :13.00   Yes:346   Yes:322
##  Mean   :50.04      Mean   :13.38
##  3rd Qu.:64.00      3rd Qu.:16.00
##  Max.   :80.00      Max.   :18.00
```



The rain cloud plot above provides data distribution, the central tendency by box plots and the jittered presentation of our raw data. When looking at the distribution of product sales volume its shape somewhat resembling a bimodal distribution (i.e. distribution with two peaks), which could potentially mean that the observations could come from two groups in terms of a currently unknown independent variable. The mean sales volume of the sample is 157.44 while the most frequent sales volume is 48 with a standard deviation of 98.45.

Based on the available data and types of measures, the actionable insights could be grouped into 3 potential action group such as changing the marketing (competitor product price, shelf placement, product price, advertising costs), growth (is_urban, is_US, population) and customer insights (education_level, representative_age, average_income). The data set will be explored along these action groups in terms of their effects on the dependent/outcome variable of interest: product_sales_volume.

Before continuing, I prepared the independent variables by standardising the continuous measures and declaring the categorical variables as factors. Standardising is important as I want to reduce the multicollinearity among the independent variables in the same model.

I selected the optimal model by using *buildmer* R package which can perform automatic backward step-wise elimination based on the change on a set criterion (here AIC values). For every action group (marketing, growth and customer insights) I first defined the maximal model including the corresponding independent variables as main effects then ran the optimal model that only contained measures that significantly improved the model fit.

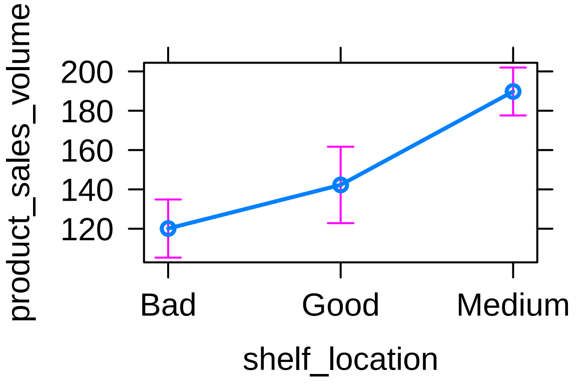
Marketing

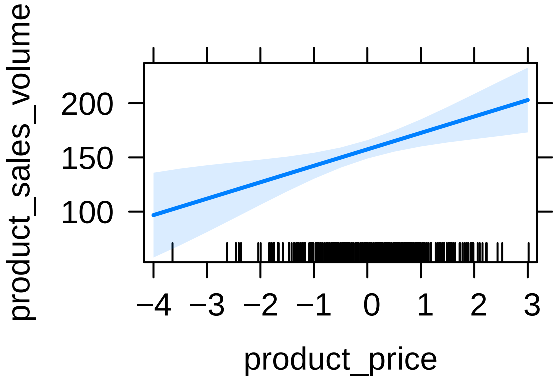
Backward step-wise elimination revealed that the optimal model was different from the maximal model of marketing as it only included competitor product price, shelf placement and product price as main effects in the model, meaning that including the measure advertising costs did not improve the model fit enough.

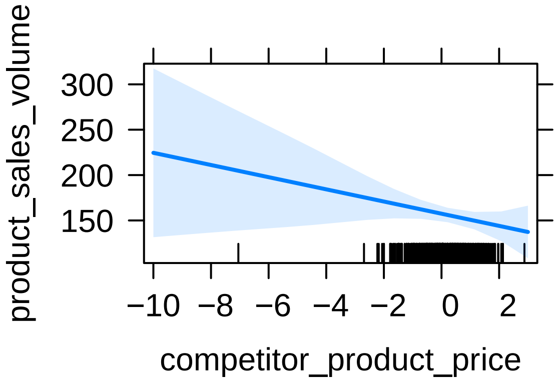
Fitting a linear regression has shown that product sales volume is significantly different across shelf types, specifically the Medium shelf location is associated with the highest levels of sales volume followed by the Good then the Bad shelf locations. A positive link between product price and sales volume was also revealed. Although, the optimal model would still include the measure competitor product price it had no significant effect on product sales volume. For more details on statistics please refer to the below summaries.

Table 1: Shelf location

shelf_location	n_obs	mean	median	stdev
Medium	221	192.05	201.0	90.96
Good	87	146.00	77.0	112.23
Bad	156	114.80	93.5	81.28







```
## MODEL INFO:
## Observations: 464
## Dependent Variable: product_sales_volume
## Type: OLS linear regression
##
## MODEL FIT:
## F(4,459) = 19.13, p = 0.00
## R² = 0.14
## Adj. R² = 0.14
##
## Standard errors: OLS
## ---------------------------------------------------------------
##                                     Est.    S.E.   t val.      p
## ------------------------------ --------- ------- -------- ------
## (Intercept)                       120.09    7.52    15.96   0.00
## shelf_locationGood                 22.18   12.59     1.76   0.08
## shelf_locationMedium               69.70    9.87     7.06   0.00
## product_price                      15.17    4.87     3.12   0.00
## competitor_product_price           -6.71    4.71    -1.42   0.16
## ---------------------------------------------------------------
```

```
## MODEL INFO:
## Observations: 464
## Dependent Variable: product_sales_volume
## Type: OLS linear regression
##
## MODEL FIT:
## F(4,459) = 19.13, p = 0.00
## R² = 0.14
## Adj. R² = 0.14
##
## Standard errors: OLS
## ---------------------------------------------------------------
##                                    Est.    S.E.   t val.      p
## ------------------------------ -------- ------- -------- ------
## (Intercept)                      142.27    9.89    14.39   0.00
## rl_shelfBad                      -22.18   12.59    -1.76   0.08
## rl_shelfMedium                    47.52   11.60     4.10   0.00
## product_price                     15.17    4.87     3.12   0.00
## competitor_product_price          -6.71    4.71    -1.42   0.16
## ---------------------------------------------------------------
```

Growth

Backward step-wise elimination revealed that the optimal model was the same as the maximal model of growth that included measures as main effects such as the size of the population, whether an area is urban or is in the US.

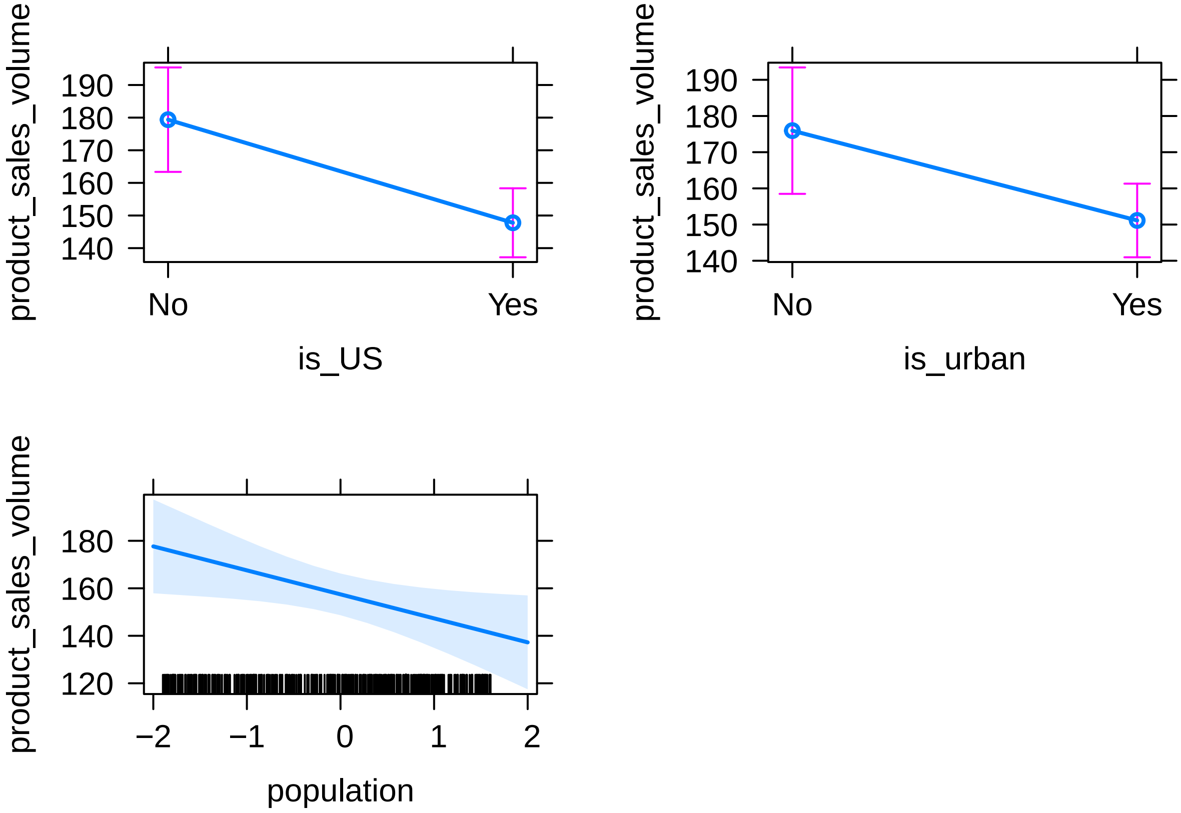
According to the results of the linear regression with the two categorical and one continuous measure, product sales volume is significantly higher in non-US areas than in the US, higher in non-urban areas than in urban areas and is negatively linked to increases in population (meaning sales volume is higher in less populated areas).

Table 2: Urban area

is_urban	n_obs	mean	median	stdev
No	118	178.86	178	98.61
Yes	346	150.14	141	97.46

Table 3: US location

is_US	n_obs	mean	median	stdev
No	142	183.06	184	84.38
Yes	322	146.15	128	102.14

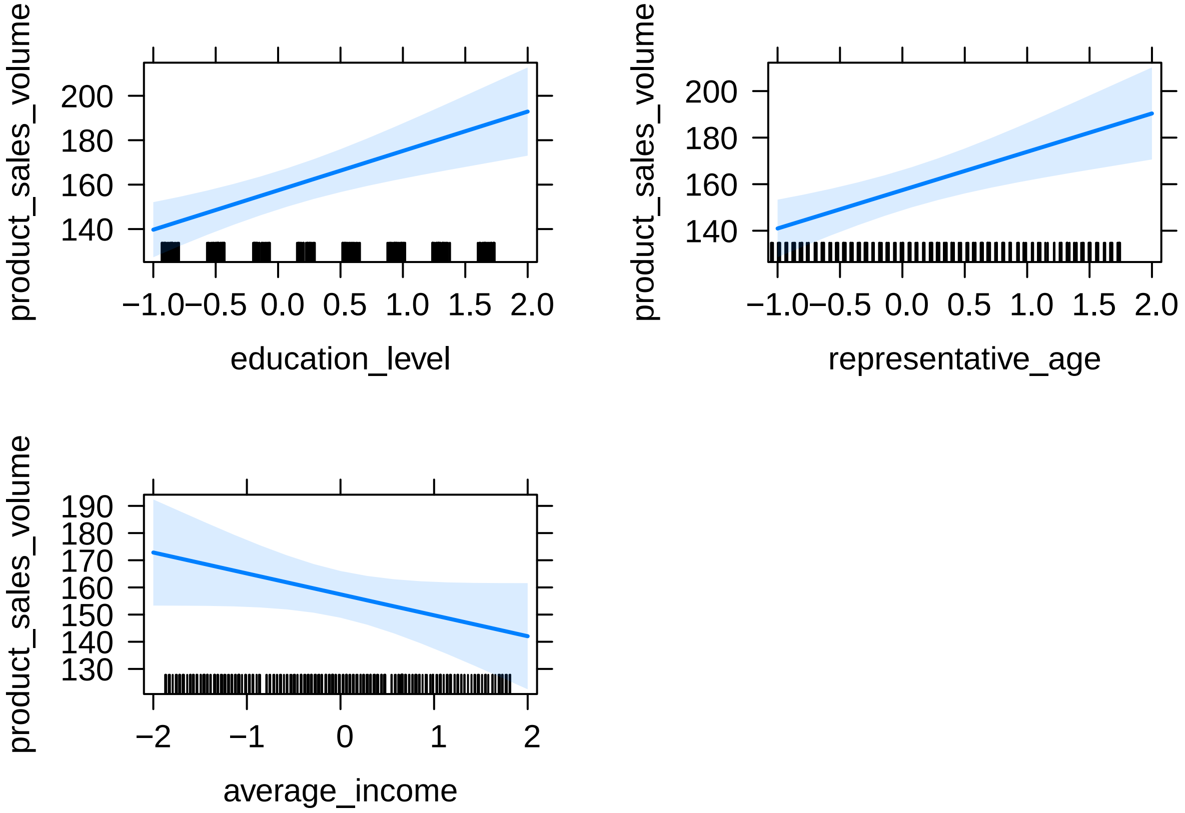


```
## MODEL INFO:
## Observations: 464
## Dependent Variable: product_sales_volume
## Type: OLS linear regression
##
## MODEL FIT:
## F(3,460) = 8.46, p = 0.00
## R² = 0.05
## Adj. R² = 0.05
##
## Standard errors: OLS
## --------------------------------------------------
##                       Est.    S.E.   t val.      p
## ----------------- -------- ------- -------- ------
## (Intercept)         197.88   10.70    18.50   0.00
## is_USYes            -31.61    9.82    -3.22   0.00
## is_urbanYes         -24.82   10.31    -2.41   0.02
## population          -10.11    4.51    -2.24   0.03
## --------------------------------------------------
```

Customer insights

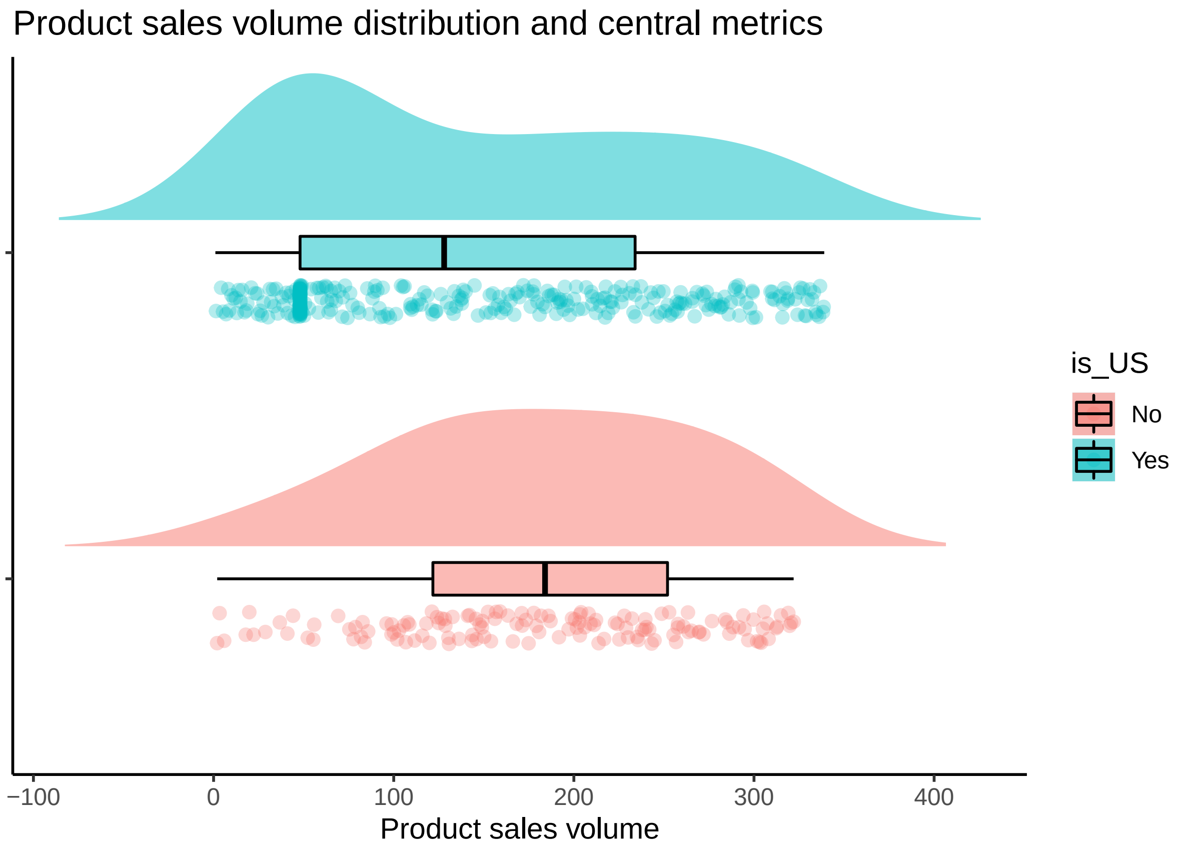
Backward step-wise elimination revealed that the optimal model was the same as the maximal model of customer insights that included education_level,representative_age,average_income as main effects.

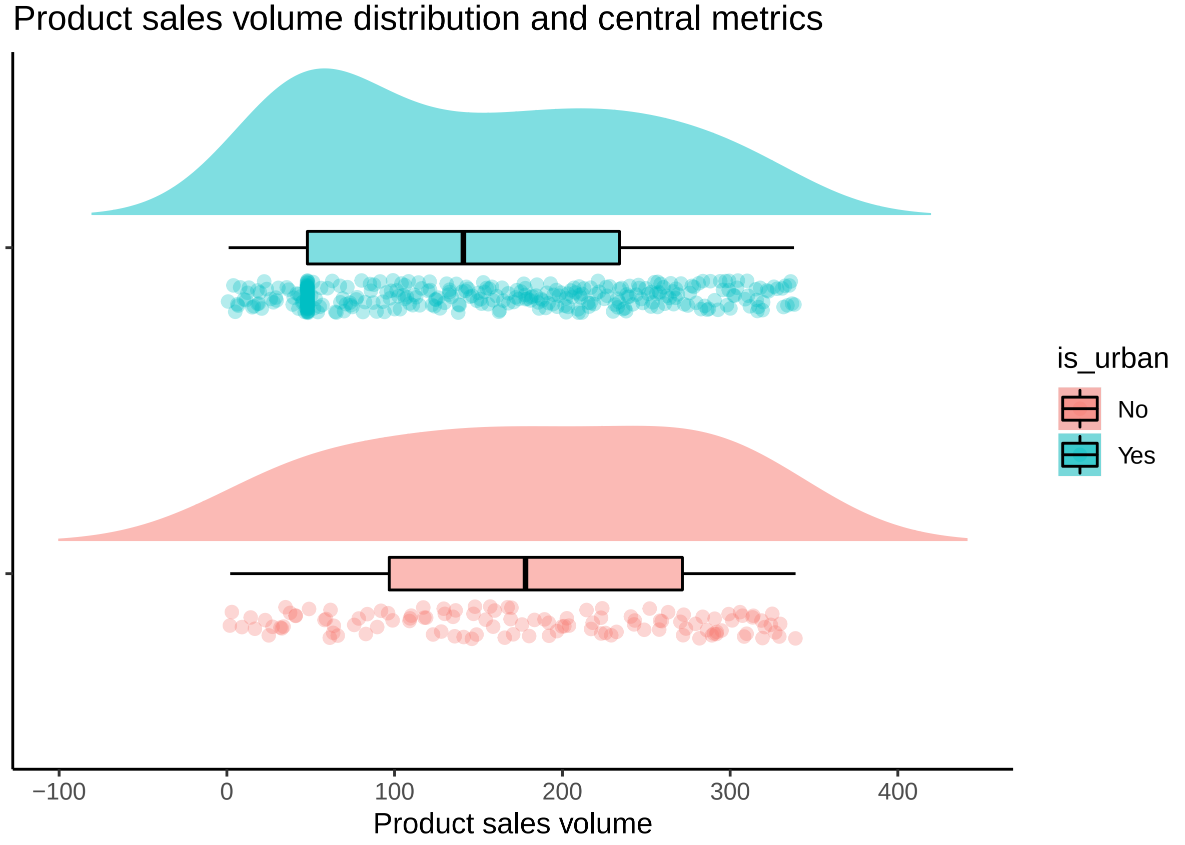
The linear regression with the three continuous independent variables suggest that product sales value increases as customers education level and age increases. Although, the optimal model would still include the measure of average income, this measure had no significant effect on product sales volume.



```
## MODEL INFO:
## Observations: 464
## Dependent Variable: product_sales_volume
## Type: OLS linear regression
##
## MODEL FIT:
## F(3,460) = 14.97, p = 0.00
## R² = 0.09
## Adj. R² = 0.08
##
## Standard errors: OLS
## --------------------------------------------------------
##                               Est.   S.E.   t val.      p
## ------------------------ -------- ------ -------- ------
## (Intercept)                 157.44   4.38    35.98   0.00
## education_level              17.73   4.56     3.89   0.00
## representative_age           16.47   4.53     3.64   0.00
## average_income               -7.70   4.46    -1.73   0.09
## --------------------------------------------------------
```

- Side note: By visually inspecting the distribution of order sales volume in the US and non-US areas and also in urban vs non-urban areas it is possible that the presence of these groups lead to a distribution with two peaks when considering the distribution of order sales volume in the whole sample. However, this hypothesis would require more knowledge about potential other variables within the two geographical groups.


Product sales volume distribution and central metrics
−100
0
100
200
300
400
Product sales volume
is_US
No
Yes


Product sales volume distribution and central metrics
−100
0
100
200
300
400
Product sales volume
is_urban
No
Yes