**Executive Summary**

This report provides an analysis and evaluation of the current and prospective customers of MINGAR Ltd. Methods of analysis include linear regression, linear mixed regression and generalized additive model as well as various plots including histogram, boxplots etc.

Result of data analyzed shows that new customers and traditional customers share similar background but with slightly difference. In particular, buyers of the newer and more affordable ‘Active’ and ‘Advance’ products come from a more widely spread age range with a slightly lower median household income. Additionally, we find out that those customers may live a community that has more population than the traditional ones.

The report finds that the prospects of the company in its current position are positive. Every time a consumer purchases products or service they display forms of buyer behavior that are influenced by many factors. Although a full comprehension of buying behavior is impossible, we can confidently predict the behavior of prospective consumers as they showed no significant difference from the current ones.

It is recommended that

* MINGAR takes immediate measures to launch and promote ‘Active’ and ‘Advance’ lines alongside its existing product range
* ‘Active’ and ‘Advance’ products adopt a more affordable and cost-efficient image to appeal customers from different age groups
* ‘Active’ and ‘Advance’ be available in discount shops as well as in traditional retail outlets

The analysis conducted has limitations, including

* Forecasting figures are not provided
* Some of the models do not have ideal statistic measures
* The current economic conditions are not provided (especially during COVID-19 era)
* There are data limitations as not enough detail is provided for the analysis i.e. how long has the customer been experiencing with the current product; how many people living under the household; has the customer ever switch product line before etc.

**Technical Report**

* Introduction

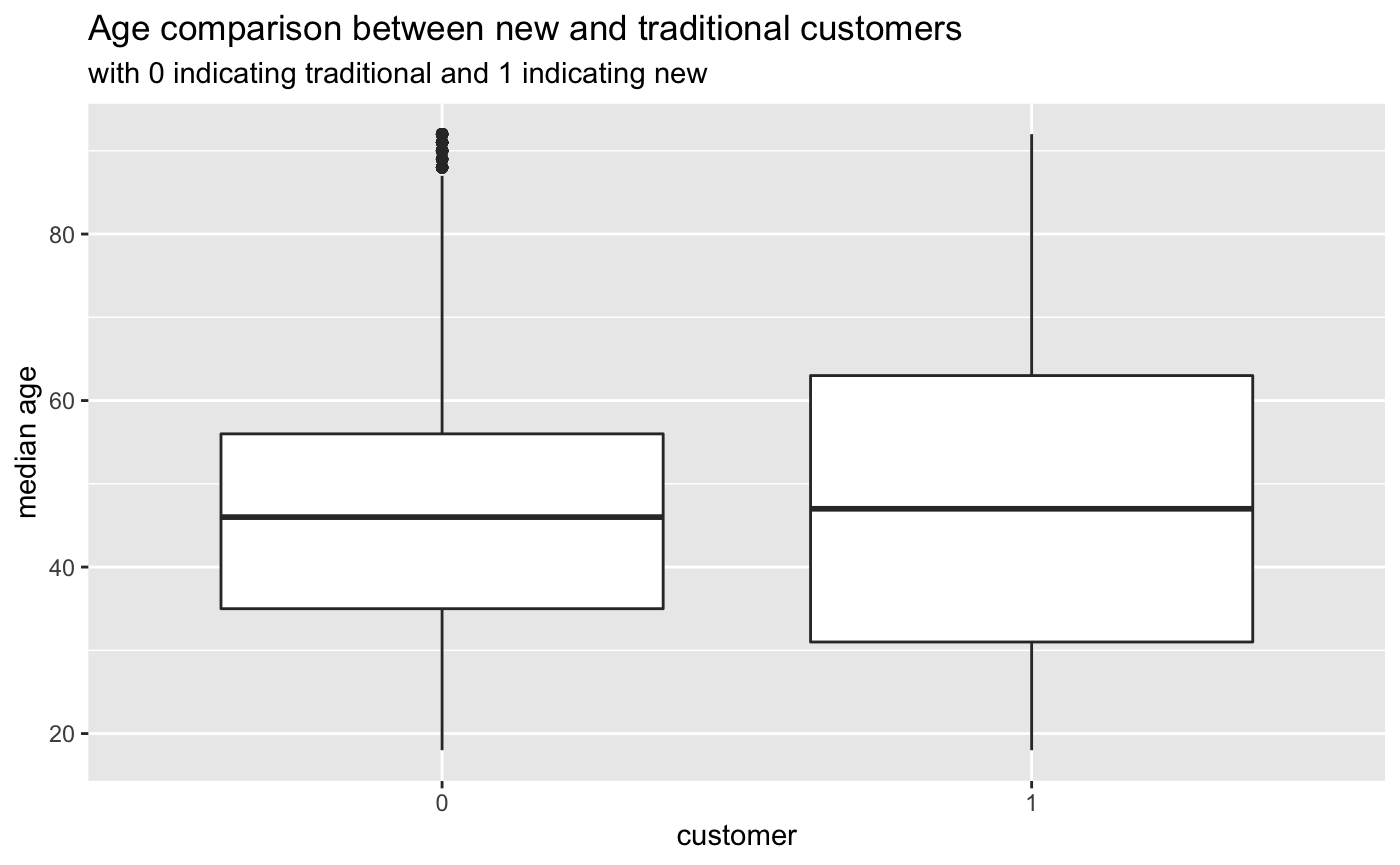
This report summarizes all of the primary statistical modeling and analysis results associated with the study of the customers of the new product lines of MINGAR LTD. The purpose of this report is to document the corresponding data modeling and inference techniques used during the subsequent statistical analyses. Additionally, this report is designed to serve as a template for describing statistical analysis technique in any future project.

The remainder of this report is organized as follows.

* The ‘Age’ subsection describes the age difference between our traditional buyers and new buyers via examining the boxplot.
* The ‘Income’ subsection summarizes the household median income difference between old and new consumers by looking into the histogram.
* The ‘Gender’ subsection produced a table containing information about the gender of our customers.
* The ‘Race’ subsection summarized the result decoded from the emoji modifier when customers using the chat and react features of the social component of the app.
* The ‘Linear Regression’ subsection examine whether there is a linear relationship between age (as responsive variable) and the classification of customers (as explanatory variable).
* The ‘Linear Mixed Regression’ subsection examine whether slope and intercept of the linear model between median household income (as responsive variable) and the classification of customers (as explanatory variable) may vary randomly depending on ‘postcode’.
* The ‘General Additive Model’ subsection uses complex functions that allow us to model non-linear relationships between population and the classification of customers. The sum of functions forms a GAM. The result is a highly flexible model which still has some of the explainability of a linear regression.
* Research Questions

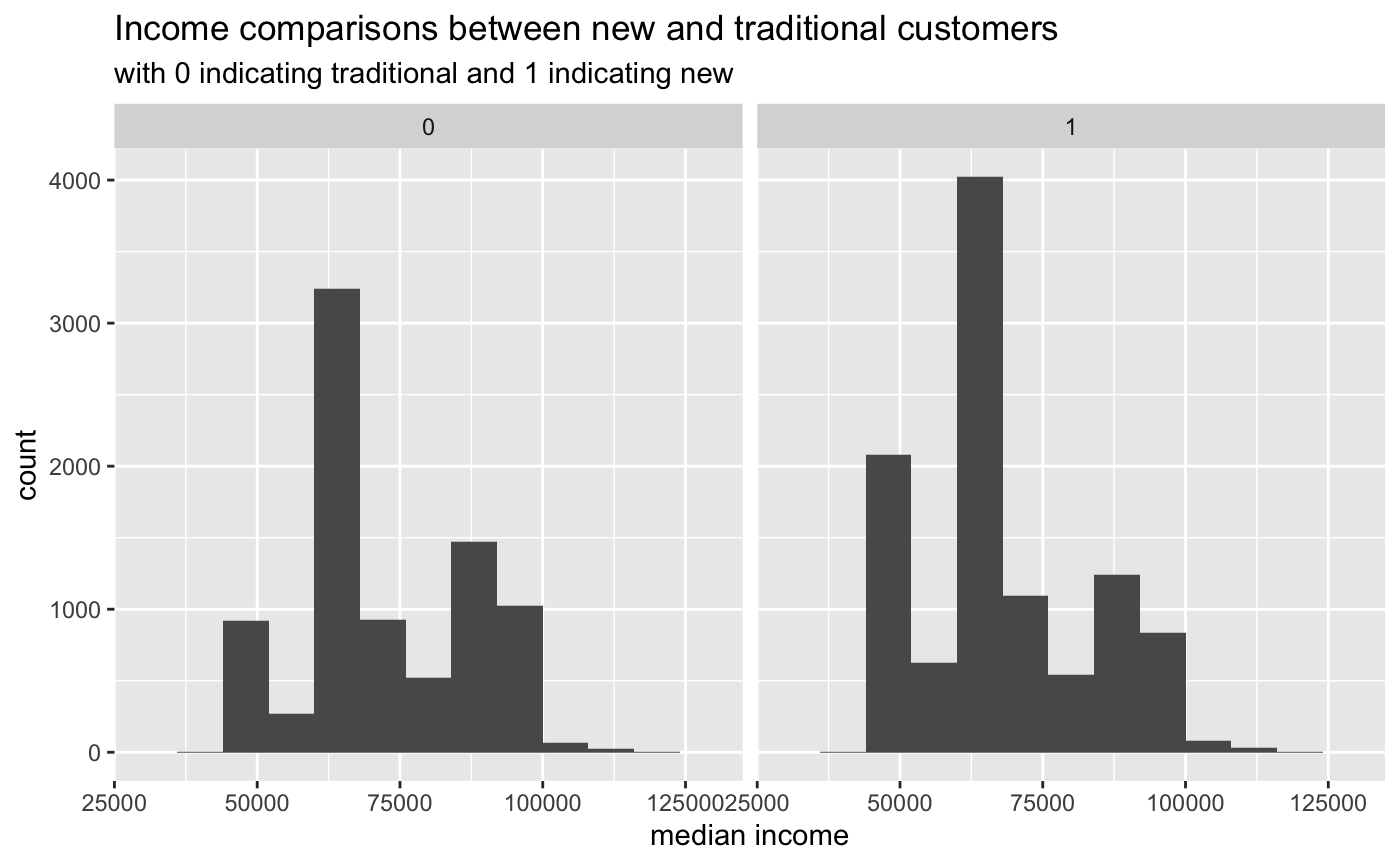
Our marketing team wants information to inform their strategy in the Canadian market. Who are our new customers? How are buyers of the newer and more affordable ‘Active’ and ‘Advance’ products different to our traditional customers?

* Age



Although having similar median ages around mid-40s, the age range of our new customers tend to be more widely spread than the traditional ones. As indicated above, the interquartile range of our new customers is significantly bigger than the traditional customers with a Q1 around early 30s and Q3 around early 60s. There are no outliers in the new customer compared to the original customers.

* + Income



Our traditional and new customers do not have a significant median income difference according to the histogram above. The median income range from both groups clusters at range 62500 and 65000.

| **new\_customer**  <fctr> | **min**  <dbl> | **Q1**  <dbl> | **median**  <dbl> | **mean**  <dbl> | **Q3**  <dbl> | **max**  <dbl> |
| --- | --- | --- | --- | --- | --- | --- |
| 0 | 41880 | 65829 | 65829 | 73168.02 | 85981 | 195570 |
| 1 | 41880 | 59881 | 65829 | 68813.94 | 77494 | 195570 |
|  |  |  |  |  |  |  |

The spread of the median income is almost identical between traditional and new customers with new customers have a slightly lower Q1, Q3 and mean value.

* + Gender

| Table 1 number of people in different sex | |
| --- | --- |
| **Gender** | **Count** |
| Female | 11081 |
| Intersex | 222 |
| Male | 7742 |

Our products tend to attract more female buyers than male buyers.

* + Race

| * Table 2 number of people in different race | |
| --- | --- |
| **Race** | **Count** |
| Dark | 2677 |
| Default | 4778 |
| Light | 3286 |
| Medium | 2606 |
| Medium~Dark | 2707 |
| Medium~Light | 2991 |

There is no significant difference of the race of customers shown in our analysis.

* + Linear Regression

mod1=lm(age~new\_customer,data= customer\_data)

summary(mod1)

Call:

lm(formula = age ~ new\_customer, data = customer\_data)

Residuals:

Min 1Q Median 3Q Max

-29.953 -13.506 -0.506 12.494 45.494

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 46.5061 0.1833 253.741 < 2e-16 \*\*\*

new\_customer1 1.4469 0.2460 5.881 4.14e-09 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.87 on 19043 degrees of freedom

Multiple R-squared: 0.001813, Adjusted R-squared: 0.001761

F-statistic: 34.59 on 1 and 19043 DF, p-value: 4.145e-09

Linear mixed model fit by REML ['lmerMod']

Formula: hhld\_median\_inc ~ new\_customer + (1 | CSDuid)

Data: customer\_data

The result is statistically significant since the p-value for this model is 4.145e-09, which is less than 0.05. The model can be interpreted using the formula: Age = 46.5061 + 1.4469(whether the customer is traditional or new).

Since the traditional customers is represented by 0 and new customer is represented by 1. We can conclude that the age of new customer is 1.4469 larger at average than our traditional customers.

* + Linear Mixed Model

mod3= lmer(hhld\_median\_inc~new\_customer+(1|CSDuid),data= customer\_data)

summary(mod3)

REML criterion at convergence: -120077.1

Scaled residuals:

Min 1Q Median 3Q Max

-4.567e-05 -1.310e-07 -7.200e-08 7.100e-08 8.750e-05

Random effects:

Groups Name Variance Std.Dev.

CSDuid (Intercept) 5.776e+06 2.403e+03

Residual 7.483e-05 8.650e-03

Number of obs: 19045, groups: CSDuid, 256

Fixed effects:

Estimate Std. Error t value

(Intercept) 7.504e+04 1.596e+02 470

new\_customer1 -9.871e-10 1.287e-04 0

Correlation of Fixed Effects:

(Intr)

new\_custmr1 0.000

optimizer (nloptwrap) convergence code: 0 (OK)

Model failed to converge with max|grad| = 69.7782 (tol = 0.002, component 1)

Model is nearly unidentifiable: very large eigenvalue

- Rescale variables?

The model used the classification of customers as fixed effect and postcode as random effect. We assume that the intercept and the slope can vary randomly among CSDuid(postcode). For fixed effects, the equation can be summarized as: Median Income = 7.504e + 04 – 9.871e-10(whether the customer is traditional or new).

Since the traditional customers is represented by 0 and new customer is represented by 1. We can conclude that the median income of new customer is 9.871e-10 smaller at average than our traditional customers regardless of postcode. The margin is relatively small so we can assume there is no significant median income difference between new and traditional customers.

* + General additive model

mod4= gam(Population~new\_customer+s(CSDuid,bs="re"),data= customer\_data, method="REML")

summary(mod4)

Family: gaussian

Link function: identity

Formula:

Population ~ new\_customer + s(CSDuid, bs = "re")

Parametric coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2166000 27692 78.217 <2e-16 \*\*\*

new\_customer1 2619 13837 0.189 0.85

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:

edf Ref.df F p-value

s(CSDuid) 0.9985 1 714.2 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) = 0.0365 Deviance explained = 3.66%

-REML = 2.8902e+05 Scale est. = 8.9078e+11 n = 19045

The result of this model is not idea since the R squared is 0.0365. However, we can interpret that new customers comes from a community with more population than our traditional customers.

**Discussion**

Based on our analysis, we found out that new customers share similar backgrounds with the original customers. It is recommended that MINGAR takes actions to launch and promote ‘Active’ and ‘Advance’ lines alongside its existing product range because the new products can benefit from current customer base. Additionally, adopting a more affordable and cost-efficient image to appeal customers from different age groups would be beneficial for the new line of products since new customers tend to have more age variations with slightly lower median household income.