Customer Segmentation

Ellen Deng

The Data

The dataset we used includes:

- 373k data points and 18 variables
- customer address, preferences, marketing channel, first order date and revenue, first 90 days revenue and days to conversion

```
Observations: 373,855
Variables: 18
$ user_id                   <dbl> 11109, 23640, 80970, 110985, 131766, 227328, 228624, 258420, 272187, 277149, 396624, 424749, 439881…
$ customer_type             <fct> lion, lion, lion, lion, lion, lion, lion, lion, lion, lion, lion, lion, lion, lion, pigeon, lion, l…
$ first_order_type          <fct> one-time, one-time, one-time, one-time, one-time, one-time, one-time, subscription, one-time, subsc…
$ region                    <fct> West, Midwest, Northeast, Midwest, South, West, West, Midwest, South, Midwest, Midwest, Northeast, …
$ subregion                 <fct> Pacific, East North Central, Middle Atlantic, East North Central, South Atlantic, Pacific, Pacific,…
$ state                     <fct> CA, IN, NJ, OH, FL, CA, CA, IL, TX, OH, MI, NJ, MD, PA, NC, AZ, CA, CA, CA, CT, TX, CA, NJ, KS, VT,…
$ city                      <fct> COLTON, INDIANAPOLIS, BORDENTOWN, SPRINGFIELD, NAPLES, MURRIETA, EMERYVILLE, OAK PARK, EL PASO, LOR…
$ post_code                 <int> 92324, 46241, 8505, 45505, 34119, 92562, 94608, 60302, 79924, 44053, 48103, 8035, 20769, 15232, 276…
$ product_preference        <fct> monitors, not specified, not specified, monitors, earphones, not specified, not specified, monitors…
$ food_preference           <fct> pizza and pasta, not specified, not specified, pizza, pasta, pizza, pizza and pasta, pizza and past…
$ days_to_conversion        <dbl> 261, NA, 25, 0, NA, NA, NA, 150, NA, NA, 0, 0, NA, NA, NA, NA, NA, NA, NA, 416, 47, NA, 226, NA, NA…
$ channel_credit            <fct> Paid Social, , Paid Social, Paid Social, , , , Paid Social, , , Other, Non-Paid, , , , , , , , Non-…
$ first_order_date          <fct> 2018-06-05, 2015-06-12, 2016-03-13, 2015-07-20, 2017-07-17, 2015-07-28, 2016-03-15, 2016-09-20, 201…
$ first_order_total_revenue <dbl> 24.95, 27.47, 24.95, 17.47, 44.90, 24.95, 54.90, 32.90, 29.95, 54.90, 49.90, 19.95, 42.90, 17.48, 0…
$ X90d_total_revenue        <dbl> 24.95, 57.42, 69.90, 17.47, 44.90, 24.95, 54.90, 57.85, 29.95, 54.90, 74.85, 44.90, 42.90, 17.48, 0…
$ first_order_source        <fct> Iterable, bing, facebook, facebook, facebook, bronto, www.huffingtonpost.com, facebook, bronto, bin…
$ first_order_medium        <fct> email, cpc, mobilenf, desktopnf, FB IG AN, email, (not set), desktopnf, email, organic, email, Refe…
$ predicted_total_ltv       <dbl> 40.87147, 98.58500, 34.95000, 22.52341, 194.74249, 54.90000, 44.92500, 196.60000, 14.97500, 174.650…
```

Executive Team

Customer Segmentation by Purchases

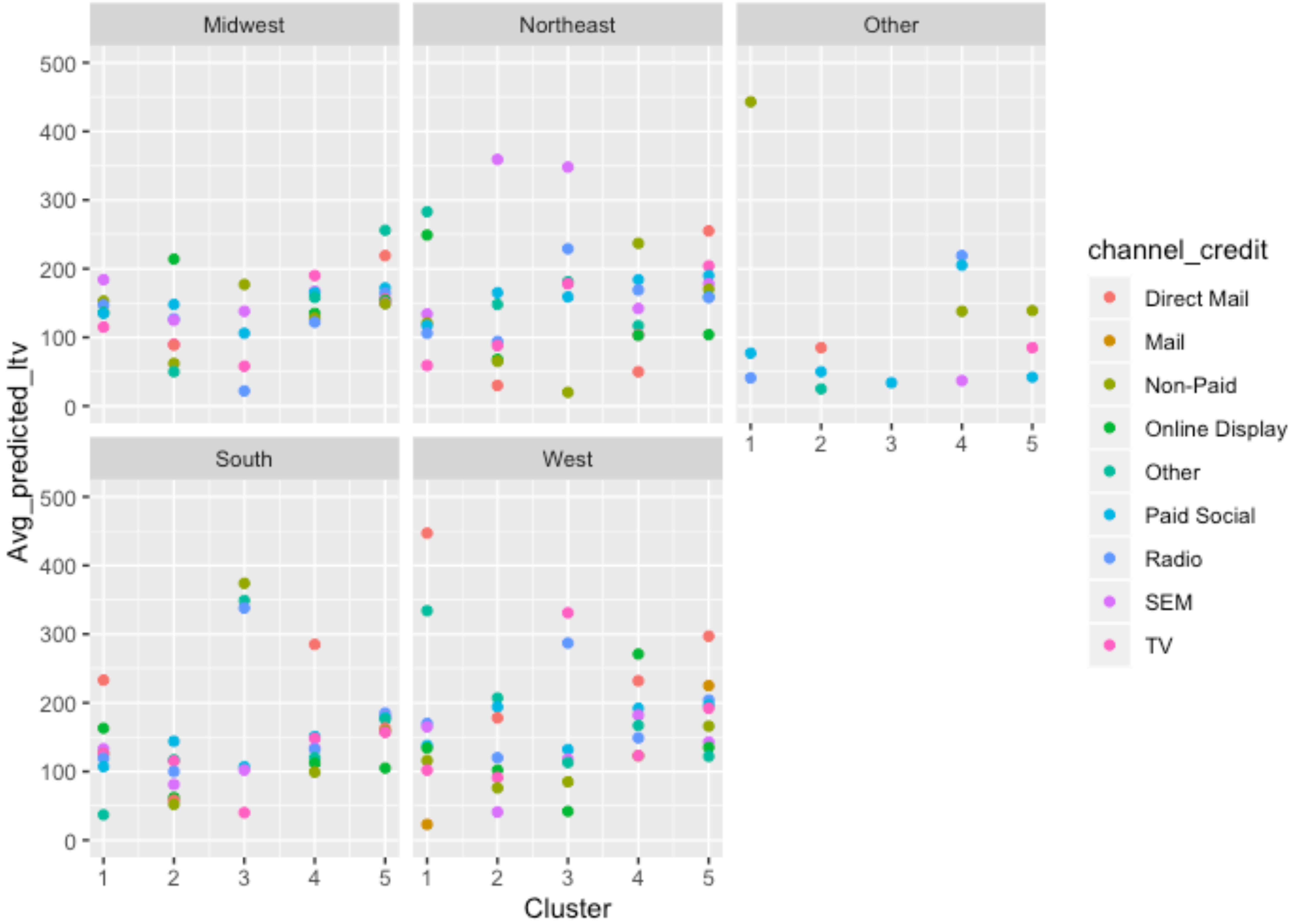
- Using K-means algorithm to conduct Cluster Analysis, we segment the customer into 5 clusters
- Although they take longer to convert, the predicted LTV is much higher
- The first_order_revenue is average among other clusters but first_90d_revenue is high compared to other clusters

```
# A tibble: 5 x 6
  Cluster Avg_predicted_ltv Avg_days_to_conversion Avg_first_order_revenue Avg_x90d_revenue
    <int>             <dbl>                  <dbl>                   <dbl>            <dbl>
1       5               170                     41                    34.1             48
2       4               149                     27                    36               48.1
3       3               128                     34                    36.2             46.1
4       1               125                     31                    33.1             42.4
5       2               119                     33                    30.8             40.1
```

Executive Team

Segmentation by Region and Marketing Channel

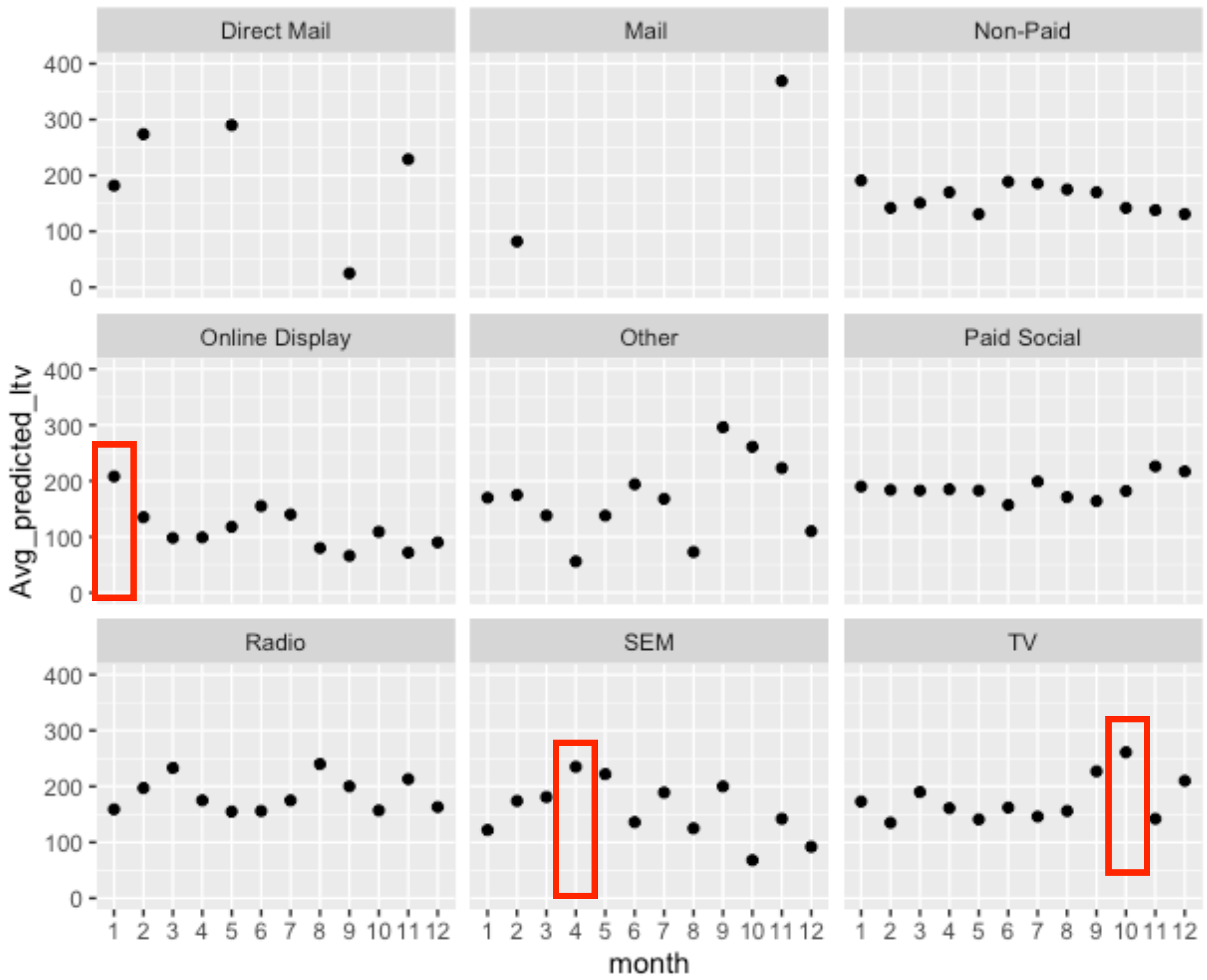
- Different marketing channels show different results in different regions
- For cluster 1 in West region, direct mail has the highest predicted LTV
- For cluster 2 in Northeast region, SEM has the highest predicted LTV
- **Target different cluster in different region with the best performing channels can acquire customers with highest predicted LTV**



Executive Team

Segmentation by Marketing Channel and First Order Month

- For Cluster 5, the marketing channel performance varies throughout the year
- TV seems to work better in October, online display seems to works better in January and SEM works the best in April
- Direct Mail and Mail seem to have high LTV but not been used very frequently
- **Prioritize different channels in different months could optimize and increase predicted LTV for cluster 5 but CAC per channel should be considered as well**



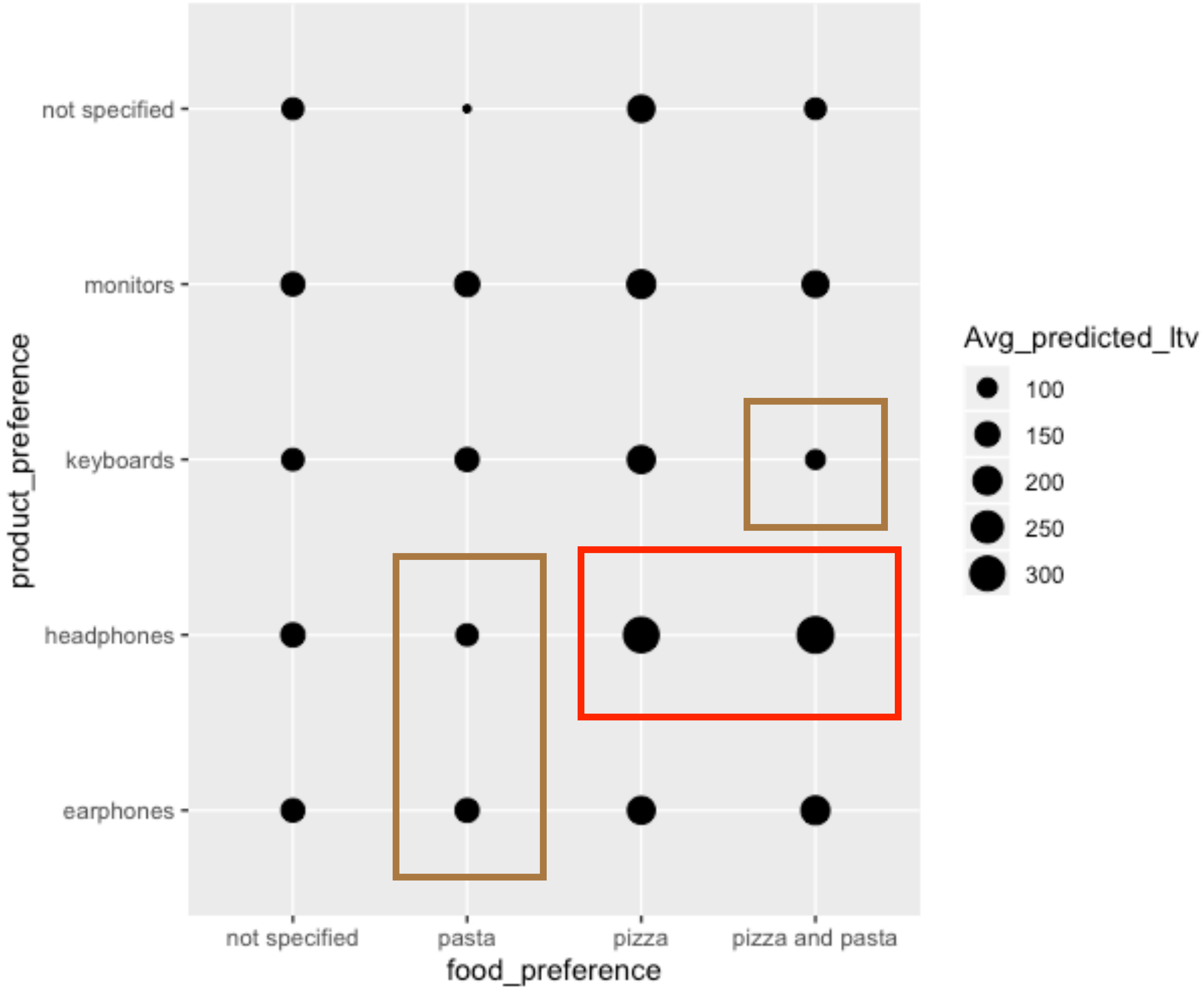
Executive Team

Segmentation by Food and Product Preference

- For Cluster 5, people who prefer headphones and pizza or pizza and pasta have higher predicted LTV

- People who prefer pasta and headphones or keyboard & pizza and pasta have low predicted LTV

- **Target people with specific preference on food an product could increase predicted LTV**



LTV Predictors

- Using liner regression model to fit the data
- first_order_total_revenue, X90d_total_revenue, first_order_type(subscription) and food_preference(pizza, pizza and pasta are the most important predictors for LTV
- Marketing channels and region are not good predictors

```
Call:
lm(formula = predicted_total_ltv ~ days_to_conversion + first_order_total_revenue +
    X90d_total_revenue + first_order_type + region + product_preference +
    food_preference + channel_credit, data = segment_customer)

Residuals:
    Min      1Q  Median      3Q     Max
-666.09  -90.23  -44.85   48.82 1277.93

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     65.47592   29.40132   2.227 0.025971 *
days_to_conversion               0.01759    0.01794   0.981 0.326838
first_order_total_revenue       -2.55824    0.12444 -20.558  < 2e-16 ***
X90d_total_revenue               3.38013    0.07968  42.421  < 2e-16 ***
first_order_typesubscription    52.99644    3.49724  15.154  < 2e-16 ***
regionNortheast                  4.17862    5.62819   0.742 0.457835
regionOther                    -35.55682   28.35497  -1.254 0.209875
regionSouth                     -8.72481    4.90127  -1.780 0.075088 .
regionWest                       5.09143    5.17254   0.984 0.324983
product_preferenceheadphones    22.22608   11.05952   2.010 0.044492 *
product_preferencekeyboards    -15.34976    7.47643  -2.053 0.040090 *
product_preferencemonitors       3.06351    5.80714   0.528 0.597829
product_preferencenot specified -20.28729   7.72224  -2.627 0.008624 **
food_preferencepasta             7.79528    5.81246   1.341 0.179908
food_preferencepizza            46.28841    5.38195   8.601  < 2e-16 ***
food_preferencepizza and pasta  20.75528    6.16960   3.364 0.000771 ***
channel_creditMail             -38.37108   88.71914  -0.433 0.665387
channel_creditNon-Paid         -35.73797   28.39595  -1.259 0.208219
channel_creditOnline Display   -39.21637   29.29832  -1.339 0.180758
channel_creditOther            -20.52288   29.24375  -0.702 0.482828
channel_creditPaid Social      -38.92782   28.16944  -1.382 0.167028
channel_creditRadio            -23.03768   28.37472  -0.812 0.416864
channel_creditSEM              -26.47359   28.76260  -0.920 0.357377
channel_creditTV               -35.73770   28.73135  -1.244 0.213581
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 168.2 on 9976 degrees of freedom
Multiple R-squared:  0.2265,    Adjusted R-squared:  0.2247
F-statistic:   127 on 23 and 9976 DF,  p-value: < 2.2e-16
```

Data Science Team

Model Fit and Selection

- From the residual plot and the QQ plot, it shows non-linearity of the data.
- It indicates linear model might not be a good fit for LTV prediction
- Due to the high dimensionality and mix data type, it might be better to use polynomial regression or tree-based method to predict LTV
- To evaluate if the model is a good fit, I will split dataset to train and test set. Then I will use cross validation to train and choose the best model. Then I will fit the model with test data and calculate MSE to evaluate the result