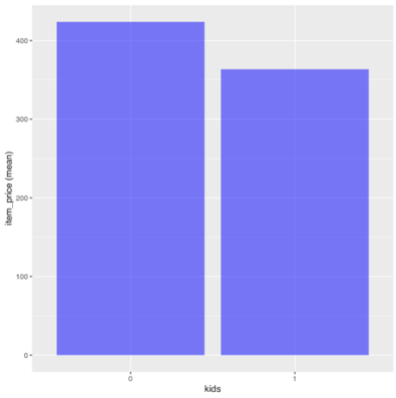
# MKT 745 Final exam Zihao Li

**Question 1**:

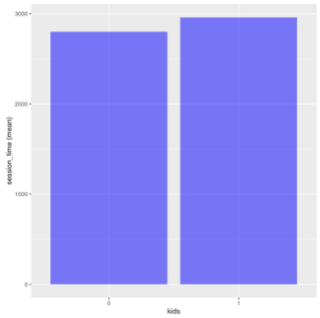
a) Model free evidence and a T-test to support or oppose his claim.




```
Pairwise mean comparisons (t-test)
Data      : CASE_FINALEXAM_CUSTOMER
Variables : kids, item_price
Samples   : independent
Confidence: 0.95
Adjustment: None

 kids    mean     n n_missing      sd     se     me
    0 423.796 2,061         0 372.959  8.215 16.111
    1 363.520   816         0 299.233 10.475 20.562

 Null hyp. Alt. hyp. diff   p.value
 0 = 1     0 > 1     60.276 < .001  ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**Model free evidence: Showed adult tend to purchase more expensive product compared to kids based on their average price of items clicked.**
**T-test: Through the T-test, we can see the p-value is less than 0.05 which means null hypothesis is rejected and the alternative hypothesis is accepted, which also showed adult tend to purchase more expensive product.**
**These support his claim.**

b) How could you address manager B's explanation? Show your logic and results. [3 points]

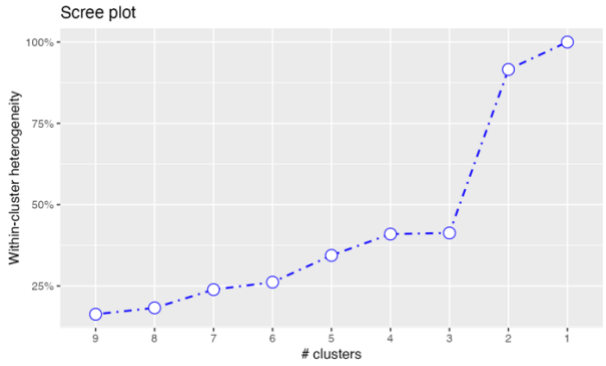



```
              coefficient std.error t.value p.value
(Intercept)       437.205    10.829  40.373  < .001 ***
session_time       -0.005     0.003  -1.782   0.075 .
kids|1            -59.519    14.627  -4.069  < .001 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-squared: 0.007,  Adjusted R-squared: 0.006
F-statistic: 10.086 df(2,2874), p.value < .001
Nr obs: 2,877
```

**The first bar chart showed adult buyers spend less time on the search for each session, but in the regression model, which session time is showing negative but the p-value which is more than 0.05 but less than 0.1, which means this could suggest a trend where longer session times might lead to slightly cheaper product selections, but this is not a strong or clear-cut finding.**

**Question 2**:

1) Employ cluster analysis to segment customers into groups.
   a. Please show the scree plot and indicate the optimal number of clusters.




**The optimal number of clusters is 3.**

   b. Show the mean statistics of all Base variables for each cluster.

```
Cluster means:
          session_cnt item_num click_num session_time item_time item_price purchase_pct kids
Cluster 1       10.44    21.80     34.54     8,241.36    300.84     370.63         0.19 0.19
Cluster 2       30.19     5.18      7.67     2,705.89    325.02     355.14         0.33 0.85
Cluster 3       13.27     5.55      7.87     2,250.68    252.92     437.99         0.12 0.00

Percentage of within cluster heterogeneity accounted for by each cluster:

Cluster 1 20.64%
Cluster 2 35.33%
Cluster 3 44.03%
```

   c. Based on the mean statistics in b, please describe the features of each cluster and define each cluster.

**Cluster 1: "Exploratory Shoppers": Mostly adult, which spend lots of time in searching and do not know what looking for and will speeding lots of time read each session material and just curious but might not purchase.**

**Lowest: session_cnt**

**Highest: item_num, click_num, session_time**

**Cluster 2: "know what they want Shoppers": looking through more item, mostly kids involve, have target in searching because item_num is low, try to find best based on their high session_cnt, which also might cause conversion rate is also highest and item_price is low.**

**Lowest: item_num, click_num, item_price**

**Highest: session_cnt, item_time, purchase_pct, kids**

**Cluster 3: "Efficient Rich Shopper": adults, conversion rate is low, in short period and willing to purchase more expensive item, but what they looks like with less compare item to competitors.**

**Lowest: session_time, item_time, purchase_pct, kids**

**Highest: item_price**

2) Based this new data, build a model to show the difference of price sensitivity across different groups and answer the following questions.
    a. Why does the sale team only select a sample of customers but not offer price discount to all customers?
    **First, they might want to see which group of people will being affect by the discount the most and by focusing on discounts for specific customer segments, companies can achieve more controlled, measurable, and effective results in their sales strategies, balancing short-term gains with long-term business health.**

    b. Show the model results, what is AIC? Please indicate which coefficients represent the difference of price sensitivity across groups.

| | OR | OR% | coefficient | std.error | z.value | p.value | |
|---|---|---|---|---|---|---|---|
| (Intercept) | | | −5.248 | 0.300 | −17.470 | < .001 | *** |
| num_click | 1.438 | 43.8% | 0.363 | 0.034 | 10.582 | < .001 | *** |
| time | 1.000 | 0.0% | 0.000 | 0.000 | 8.338 | < .001 | *** |
| item_price | 0.982 | −1.8% | −0.019 | 0.044 | −0.417 | 0.676 | |
| cluster\|2 | 9.626 | 862.6% | 2.265 | 0.302 | 7.498 | < .001 | *** |
| cluster\|3 | 2.649 | 164.9% | 0.974 | 0.315 | 3.094 | 0.002 | ** |
| item_price:cluster\|2 | 0.892 | −10.8% | −0.114 | 0.052 | −2.190 | 0.028 | * |
| item_price:cluster\|3 | 0.978 | −2.2% | −0.022 | 0.051 | −0.435 | 0.664 | |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Pseudo R-squared:0.162, Adjusted Pseudo R-squared:0.158
AUC: 0.823, Log-likelihood: −1546.02, AIC: 3108.04, BIC: 3165.722
Chi-squared: 596.692 df(7), p.value < .001
Nr obs: 10,000

**AIC : 3108.04**
**Item_price -0.019 shows the coefficient for cluster 1**
**-0.114 shows the difference of price sensitivity across groups between cluster 1 and 2**
**-0.022 shows the difference of price sensitivity across groups between cluster 1 and 3**

    c. What are the price coefficients for these groups separately?
    **Cluster 1: -0.019 (P-value is more than 0.05 so not significant)**
    **Cluster 2: -0.133**
    **Cluster 3: -0.041**

    d. Based on the feature of each cluster as you observed in Question 2-1)-c, please come up with an explanation for cluster with highest price sensitivity
    **Cluster 2 has highest price sensitivity, it might cause by they know what they looking for and through compare each similar item to get what they want which means thwy are looking for most competitive thing like cheaper in price.**

    e. Literature shows that when customers spend more time on search, they would obtain more information of product quality. In such a way, they will become less price sensitive regardless of their income level. Do you observe any evidence from price coefficients of these clusters which could support this statement? Why?
    **If only look at the time's coef, time spent has a very small, practically negligible effect.**

**But if based on mean data from cluster analysis and price coefficients, we can see when the mean session time increasing, and the price coefficient is decreasing compare between cluster 1 to 2 and 3. But this can not be and evidence but only can see this happen during this situation but can not be a proof.**

3) Please report which group is the selected customers for this campaign, and why.
**I would recommend 2nd cluster doing the campaign. Because this cluster is most sensitive to the price which will have the most affect. Which mean we can cost least but increase sales the most.**

**Question 3**: [60 points]

1) Please fill the table by calculating the performance metrics and cost metrics.

| Campaign | keywords | target | period | CTR | Conversion rate | CPM | CPC | CPA | ROI |
|---|---|---|---|---|---|---|---|---|---|
| Google search ads | kid toy | NA | May 2-May 6 2018 | 0.90% | 14.50% | $96.2983 | $10.70 | $73.79 | 1.21 |
| Google search ads | kid gift | mobile user | June 16-June 21 2018 | 1.20% | 12.00% | $169.1985 | $14.10 | $117.51 | 1.23 |
| Facebook ads | discount | users with kids | Mar 1-Mar 19 2018 | 0.10% | 6.88% | $5.6994 | $5.70 | $82.81 | 1.28 |
| Facebook ads | NA | young male | Jul 21-Aug 2 2018 | 0.20% | 8.29% | $8.3996 | $4.20 | $50.66 | 4.20 |

2) If the purpose of the campaigns is to increase the sales in official website, which cost metric would you use to evaluate these campaigns? And based on this cost metric, which campaign has the best performance?
**I would say CPA and ROI because CPA and ROI are most directly tied to evaluating the effectiveness of campaigns in terms of actual sales performance. And based on cost, Facebook ads to young male will cost least but show the best performance.**

3) Which performance metric is related to this cost metric? Write down the formula to show such relationship and explain the implication of this formula.
**Conversion rate is related to CPA.**
**CPA = CPC/conversion rate. Which means it based on conversion rate to calculate how much is will cost to make sure it will one customer to have purchase action.**

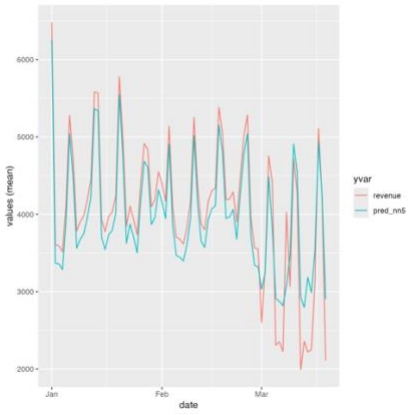
4) Please briefly describe the logic on why and how we use counterfactual analysis to evaluate the impact of online campaigns on offline sales.
**Because it tries to evaluate the impact of online campaigns on offline sales by comparing what happened with what would have happened in the absence of the online campaign. Like try to control the most and find the different under only one different situation.**
**First is build a prediction model of outcome based on historic data and decide the optimal size of NNM also save RMSE and MAE of all data for the final model. Then evaluate the performance of prediction model. Do the counterfactual analysis by predicting the outcome**

**during campaign period and measure the campaign effect. Lastly do the Heterogeneity analysis like: Temporal variation, Cross section variation.**

5) Report MAE for training and test data in the following table.

| | Training | Test |
|---|---|---|
| Size=1 | 597.193 | 658.901 |
| Size=3 | 418.789 | 445.184 |
| Size=5 | 364.829 | 387.965 |
| Size=7 | 374.489 | 395.995 |
| Size=9 | 376.181 | 405.499 |

6) What is the most important assumption for counterfactual analysis? Please use two evidence to show whether this assumption holds for the model with optimal size.
**The most important assumption in counterfactual analysis is the parallel trends assumption. This assumption posits that, in the absence of the intervention the trends in outcomes for both the treated and control groups would have been the same over time.**
**Pre-Campaign Trend Comparison: Compare sales between treatment and test group. If the trends were similar before the campaign, this supports the parallel trends assumption.**
**Post-Campaign Deviation from Trends: After establishing a baseline pre-campaign trend, check whether the post-campaign sales deviate from this expected trend only in the treated group and not in the control group.**
**MAE% = 10.1% which is less than 20 percent.**




**We can see before March, there is parallel trend but after than the predict and actual is showing different.**

7) By conducting counterfactual analysis with model of optimal size, report the average effect of Facebook Ads on the sales of shop stores and its 95% prediction interval. Compared with the results in 1), please conclude the effects of Facebook Ads campaign on sales in online and offline stores.
**Mean for diff in campaign period is 3512.711.**
**Mean for diff out of campaign period is 4106.807.**
**Mean for diff is 3962.106.**
**Lower: 2337.743 upper = 4687.679**

**During the campaign period, the average revenue or sales is lower than compared to before campaign period which from Jan to Feb.**

8) The team is also wondering whether the effect would be different in terms of stores and timing. Please use the variables in the out-sample data to conduct two heterogeneity analyses to show and conclude how the effect varies across store location and weekday respectively. Please show how confident you believe your conclusions? Please come up with explanations for your conclusions.

| | | Function |
|---|---|---|
| weekday | variable | mean |
| All | All | All |
| 1 | dif | 4,211.560 |
| 2 | dif | 2,915.812 |
| 3 | dif | 2,835.088 |
| 4 | dif | 3,002.110 |
| 5 | dif | 3,032.551 |
| 6 | dif | 3,418.649 |
| 7 | dif | 4,777.135 |

```
weekend      mean  n n_missing        sd      se        me
      0 3,059.648 52         0 1,529.804 212.146   425.900
      1 4,494.348 24         0 2,487.983 507.857 1,050.583

Null hyp. Alt. hyp. diff      p.value
0 = 1     0 < 1     -1434.699 0.007   **

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**It has most campaign effect on Saturday and least campaign effect on Monday.**

**Or It has most campaign effect on weekend and least campaign effect on weekday.**

| | | Function |
|---|---|---|
| citytier | variable | mean |
| All | All | All |
| large | dif | 4,349.511 |
| small | dif | 2,675.912 |

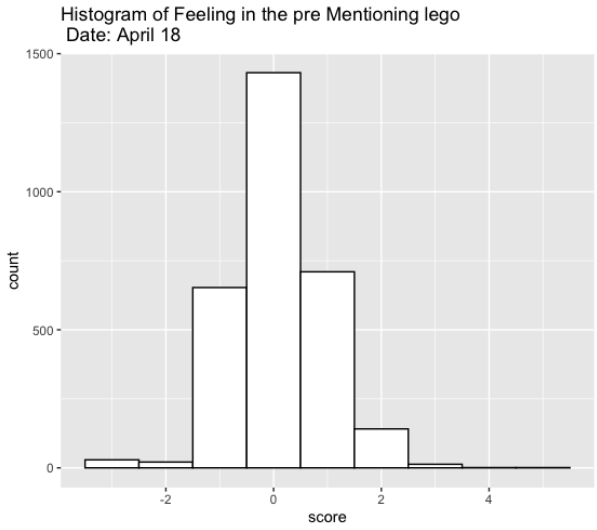
```
citytier      mean  n n_missing        sd      se      me
   large 4,349.511 38         0 2,377.352 385.657 781.416
   small 2,675.912 38         0   947.153 153.649 311.322

Null hyp.       Alt. hyp.       diff   p.value
large = small   large > small   1673.6 < .001  ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

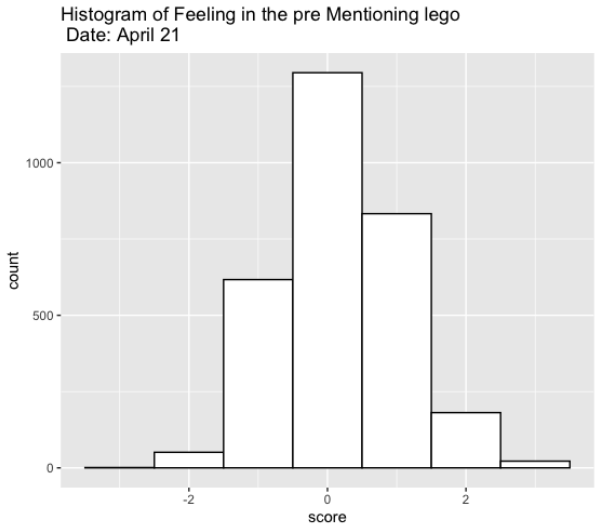
**It has most campaign effect on large city tier and least campaign effect on small city tier.**

**Based on the t-test for how the effect varies across store location and weekday respectively, which both showed the same result as what I assumed.**

**Question 4**:

*Question*: Please show the histogram of people's feeling as well as the average feelings for twitters mentioning "Lego" on April 18 and April 21 respectively. Compare how people feel change after the announcement on April 20. [5 points]




**4-18: 0.2346833**

**4-21: 0.2187333**

**After the announcement, the neutral feeling decreasing, but have bit more positive feeling, but the negative on -2 also increased a bit. But in generally, the average feeling decreased after the announcement.**