Group 7: Jialiang Cui, Lara Daoud, Siow Meng Low, Rémi Sabonnadiere

Digital Marketing Analytics

Group Homework 2

Question 1: A/B Testing

Context

A landing page of a website can contain important information that the creators of the website would want visitors to see. Because the landing page is an entry point to the rest of the website, how it is designed and how visitors navigate on it is of great importance. Therefore, an A/B test on multiple aspects of the design of the landing page can be carried out to optimise its design. This can include how the navigation buttons are designed on the website.

The original design of the landing page includes top navigation bar which can be used to navigate to other pages. It is visible at the top as soon as the page loads and visitors can navigate to other pages straightaway without browsing the whole landing page. However, in some cases the creator of the website would want visitors to browse the landing page before deciding on which page to visit next, as navigation buttons placed further down the page could include a more detailed description of what each page would include, making it more likely that the page the visitor clicked on is the page they are looking for. From the creator's point of view, this gives a more accurate view of what visitors are looking for when visiting the page, as opposed to having visitors navigate to multiple pages and not knowing the page they are interested in.

For the aforementioned reason, an A/B test was run to check if the top navigation bar at the landing page causes the visitors to skip browsing the landing page. The success criterion is defined as clicking on the buttons placed on the landing page (and not using the top navigation bar).

Null and Alternative Hypotheses

Since we define the success criterion as clicking on the buttons placed on the landing page, we can use Click-Through Rate (CTR) of the buttons as the metric:

$$CTR = \frac{Number\ of\ Clicks\ on\ the\ buttons}{Number\ of\ Impressions\ (i.e. Number\ of\ Visits\ to\ Landing\ Page)}$$

We then define the hypotheses as:

- Null Hypothesis: Having top navigation bar (Scenario A) does not decrease the chances of visitors scrolling down, browsing and clicking on buttons on the landing page.

$$H_0: CTR_A = CTR_B$$

- Alternative Hypothesis: Having top navigation bar (Scenario A) affects the number of visitors scrolling down, browsing and clicking on buttons on the landing page (Two-Tail Test).

$$H_1: CTR_A \neq CTR_B$$

Experiment Details

To conduct A/B testing, we would need two scenarios (scenario A and scenario B) and experiment if there are any differences in visitor behaviours.

Two Variants of Websites

Version A: this is the original design of the landing page. Top navigation bar are placed at the top of the page which can be used to navigate to other pages.

Version B: top navigation bar is removed from the landing page in this version. Only the title of the page is visible at the top when the page loads. The buttons are placed further down the pages. These buttons are kept in the exact same design as the original version (Version A) to ensure that only the effect of top navigation bar is measured with this test.

The different versions of websites are shown below:

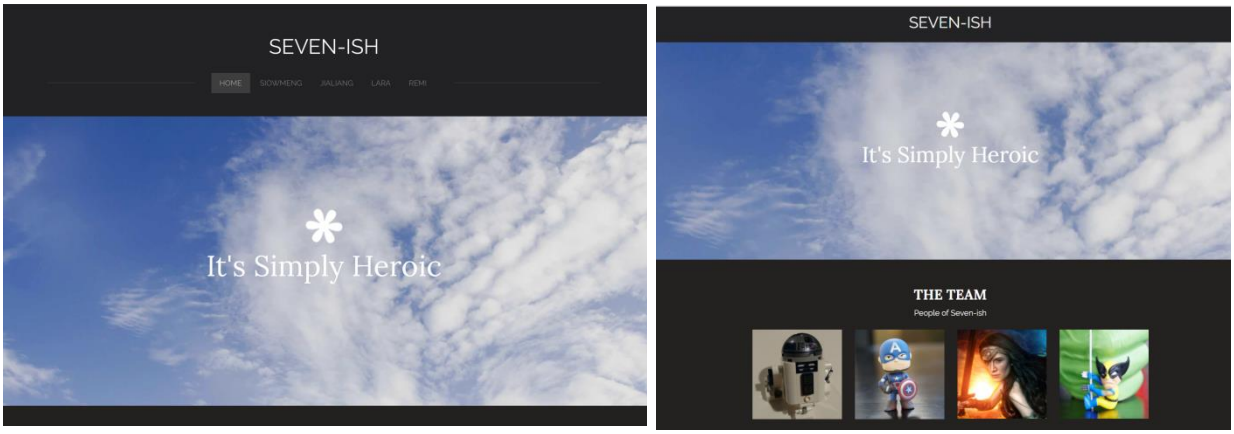


Figure 1 Two Versions of Home Page (Left: Version A, Right: Version B)

Percentage of Visitors Targeted

100% of the visitors of the website were included in the experiment and received either version A or version B of the website.

Distribution of Visitors to Version A & B

Both versions of the website were given equal weights in the experiment. Approximately 50% of visitors were shown version A of the website and the rests were shown version B.

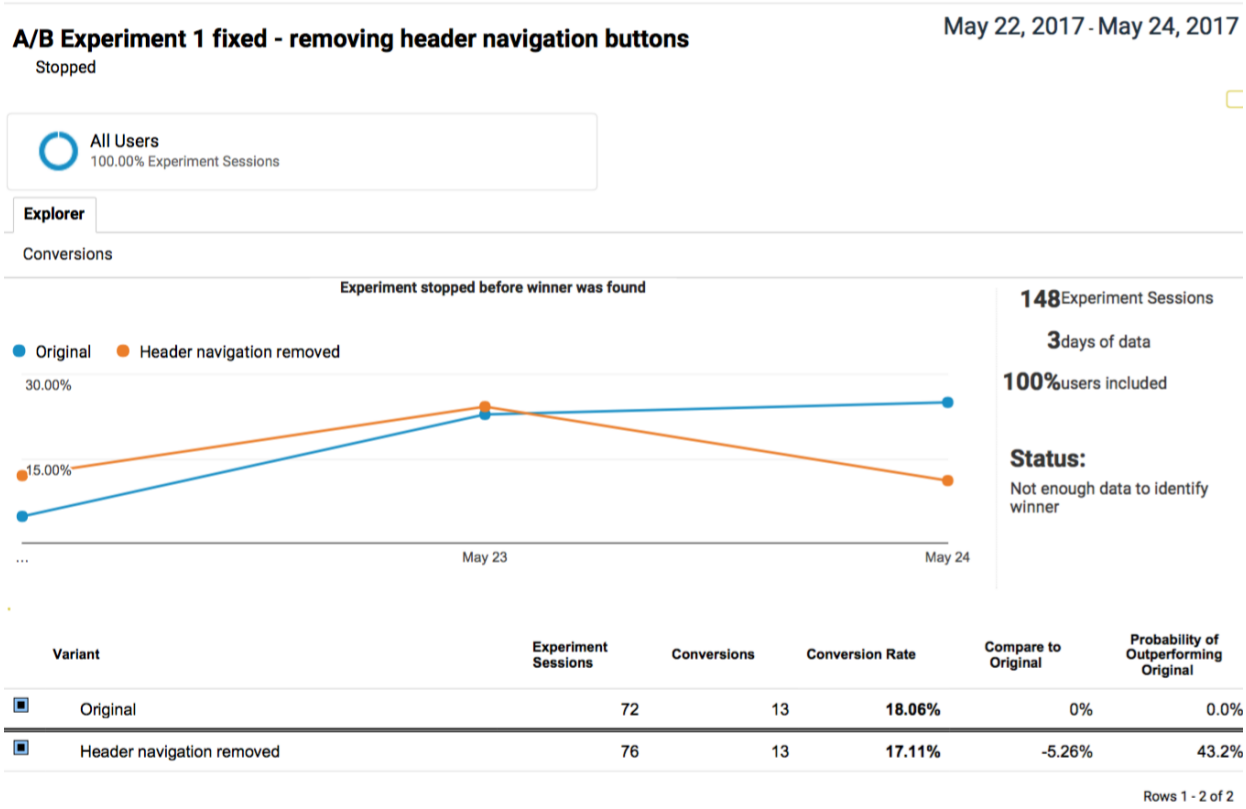
Performance Metric

In order to decide which version is better, a goal is needed to be used as a metric to compare the performance of both versions. Since the aim of the test is to measure whether visitors scroll through the page and click on the buttons placed at the landing page, a success is defined as a click on buttons placed further down the page, and not using the top navigation bar to navigate to other pages.

Based on this reasoning, the better performing version is therefore the version that achieved a higher Click-Through Rate on buttons placed further down the landing page.

Experimentation Results on Google Optimize

The final results of the experiment (on Google Optimize) are shown below:



Variant	Experiment Sessions	Conversions	Conversion Rate	Compare to Original	Probability of Outperforming Original
Original	72	13	**18.06%**	0%	0.0%
Header navigation removed	76	13	**17.11%**	-5.26%	43.2%

Rows 1 - 2 of 2

Figure 2 Results from Google Optimize

The results show that even with the top navigation bar (version A – blue curve in the above screenshot), the original design still received more clicks on buttons placed furthers down the landing page than the version without the top navigation bar (version B – orange curve in the above screenshot). The original version (version A) achieves click-through rate of 18.06% and the version with top navigation bar removed (version B) has 17.11%.

The CTR of version B is slightly lower and we will test whether this difference is statistically significant in the next section

Hypothesis Testing Result

The hypotheses formulated need to be tested statistically to determine whether the Click-Through Rates of the buttons placed further down the landing page were equal in both versions to check for consistency with Google Optimize results.

Two sample T-test

As we are comparing the CTRs of two versions, a two-sample T-test is appropriate to test the proportions of button clicks relative to the number of visits. T-test is preferred over Z-test because we have relatively small sample size (around 70+ observations) and the population standard deviations are unknown. Therefore, we use the two-sample T-test since it is more suitable to our dataset.

The result of the two-sample T-test is tabulated in Table 1 below:

	Version A	*Version B*
Mean	0.180555556	0.171052632
Variance	0.150039124	0.143684211
Observations	72	76
Hypothesized Mean Difference	0	
df	145	
t Stat	0.150736447	
P(T<=t) one-tail	0.440196557	
t Critical one-tail	1.655430251	
P(T<=t) two-tail	0.880393114	
t Critical two-tail	1.976459563	

Table 1 Hypothesis Testing 1

The two-tail p-value of T-test is very high in the above table, around 0.88, hence the null hypothesis that click through rates in both versions is equal could not be rejected. Since there is statistically insignificant difference between the two CTRs, we conclude that that having top navigation bar does not affect the behaviours visitors in scrolling down, reading the landing page and clicking on the buttons placed at the landing page. Therefore, we would continue to adopt the original design (top navigation bar is present) since there is no positive effects in removing it.

Part 2: Statistical Testing

RG is aiming to test the efficacy of their ads business. To do this, they pick 30,000 restaurants and divide them into three treatment groups.

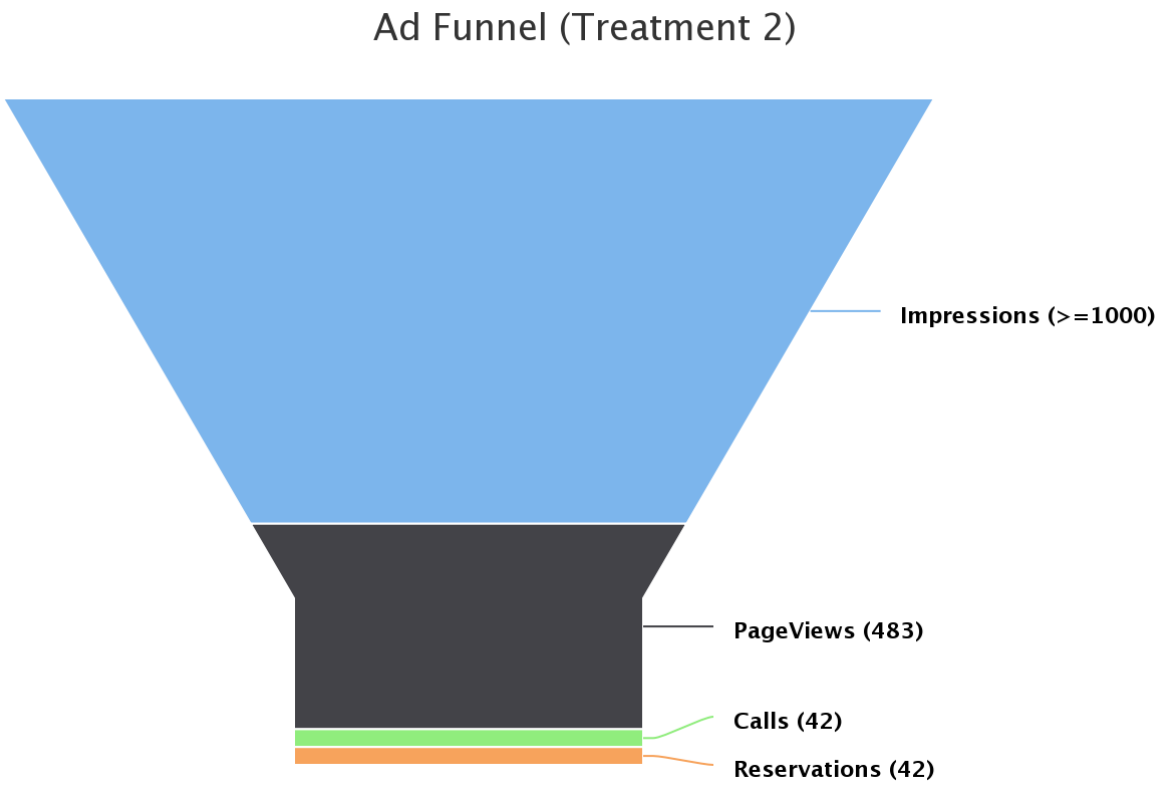
Treatment 0	Control group without ads
Treatment 1	Current ads (trigged by type of cuisine within a 0.5-mile radius of user search)
Treatment 2	Alternative ad design (trigged by a specific restaurant search and shows restaurants with similar rating and hours)

Table 2 List of Three Treatment Groups

Hypotheses Formulation

Each restaurant subscribing RG's ads service (e.g. in first and second treatment group) is guaranteed to show ads for at least 1000 times per month. In this case, the actual impressions data (how many times ads are shown) are not available. As a result, metrics like Click-Through Rate (CTR) are not available.

Below diagram shows the funnel of treatment group 2, displaying average number of impressions, page views and reservations.



** The figures for actual impressions are not available but for the restaurants with ads, it's guaranteed to be more than 1000*

Since we cannot use CTR for the analysis, to test whether current ads and alternative ad design are effective in bringing in relevant diners, conversion rate is the major metric here. Conversion rates are calculated as follows:

$$\text{Conversion Rate} = \frac{\text{reservations}}{\text{page views}}$$

Three two-sample T-tests are used to test whether the conversion rates are different across three groups. The three hypotheses are as follows:

1	Null Hypothesis Treatment 0 has the same mean conversion rate as Treatment 1 Alternative Hypothesis Treatment 0 has different mean conversion rate than Treatment 1
2	Null Hypothesis Treatment 0 has the same mean conversion rate as Treatment 2 Alternative Hypothesis Treatment 0 has different mean conversion rate than Treatment 2
3	Null Hypothesis Treatment 1 has the same mean conversion rate as Treatment 2 Alternative Hypothesis Treatment 1 has different mean conversion rate than Treatment 2

Table 3 Three Sets of Hypothesis Testing

Results of Two-Sample T-Tests

Table 4 shows the two-sample T-test result comparing the mean conversion rates of the control group (Treatment 0) and first treatment group (Treatment 1).

It can be seen that the control group has a higher mean conversion rate than the first treatment group and two-tail P-value is essentially zero. This indicates that the difference in mean conversion rate is extremely significant statistically. Since the control group has a statistically higher mean than Treatment 1, using the current ads actually worsens the conversion rate, compared to not using ads at all.

	Treatment 0	*Treatment 1*
Mean	0.087376226	0.071681758
Variance	0.000449626	0.000207241
Observations	10000	10000
Hypothesized Mean Difference	0	
df	17601	
t Stat	61.23615459	
P(T<=t) one-tail	0	
t Critical one-tail	1.644940204	
P(T<=t) two-tail	0	
t Critical two-tail	1.960098774	

Table 4 Hypothesis Testing 1

Next, in Table 5, we would like to test the effectiveness of alternative ad design (against not using ad), hence we conduct a T-test to test the difference in mean conversion rates of control group (Treatment 0) and second Treatment group (Treatment 2). From Table 5, it can be seen that Treatment 2 has a higher mean than control group and this difference is statistically very significant (two-tail P-value is essentially zero). Therefore, we conclude that Treatment 2 is statistically more effective in bringing in diners than not having ads at all.

	Treatment 0	*Treatment 2*
Mean	0.087376226	0.093575719
Variance	0.000449626	0.000552934
Observations	10000	10000
Hypothesized Mean Difference	0	
df	19788	
t Stat	-19.57947279	
P(T<=t) one-tail	7.30449E-85	
t Critical one-tail	1.644930635	
P(T<=t) two-tail	1.4609E-84	
t Critical two-tail	1.960083876	

Table 5 Hypothesis Testing 2

The last T-test is to compare the difference in mean conversion rate between Treatment 1 and Treatment 2. From Table 6, it can be seen that Treatment 2 has a higher mean conversion rate and it is again statistically very significant (two-tail P-value is essentially zero). Therefore, we conclude that the alternative ad design is more effective in bringing in diners than the current ads.

	Treatment 1	*Treatment 2*
Mean	0.071681758	0.093575719
Variance	0.000207241	0.000552934
Observations	10000	10000
Hypothesized Mean Difference	0	
df	16571	
t Stat	-79.40859186	
P(T<=t) one-tail	0	
t Critical one-tail	1.644945586	
P(T<=t) two-tail	0	
t Critical two-tail	1.960107153	

Table 6 Hypothesis Testing 3

Table 7 shows the summary statistics of the three groups (average page views, reservations and mean conversion rates).

	Average Page Views	Average Reservations	Mean Conversion Rate
Treatment 0	419.7794	33.9604	8.74%
Treatment 1	501.1908	34.0212	7.17%
Treatment 2	483.211	41.6805	9.36%

Table 7 Summary of three Treatment Groups

The summary table shows Treatment 2 has more reservations and the three t-tests have confirmed it has statistically higher conversion rate than the other 2 groups. The result indicates new ad design is more efficient and effective; therefore, new ad design is recommended to RG.

Mean Conversion Rates of Different Restaurant Types

In these 30,000 restaurants, 18,000 are independent restaurants and 12,000 are chain restaurants. To rule out that the significant difference in mean conversion rate is due to the restaurant types (rather than ad design), we also conduct the same two-sample T-tests (i.e. Treatment 0 vs Treatment 1, Treatment 0 vs Treatment 2, and Treatment 1 vs Treatment 2) on independent restaurants only and on chain restaurants only.

In both of these two cases, the differences in mean conversion rates across the three groups are all statistically significant. This is consistent with our earlier T-test result on the whole dataset hence the hypothesis testing outputs are not shown here.

Table 8 shows the summary statistics of three groups, break down by restaurant type. Consistent with our earlier results, the alternative ad design does not improve too many page views than current ad design; but alternative ad design improves the conversion rate a lot. Current ads are not as effective in converting page views to reservations. Based on RG's business model that restaurants subscribe to the ad service and pay for monthly subscription fee, it is concluded that current ad design is not as effective as alternative ad design. The new ad design also achieves higher conversion rate than having no ad at all, hence the new ad design is recommended to RG.

	Mean Conversion Rate (Chain)	Mean Conversion Rate (Independent)	Page Views (Chain)	Page Views (Independent)
Treatment 0	6.68%	10.11%	599.5712	299.9182
Treatment 1	5.83%	8.06%	690.3967	375.0535
Treatment 2	6.96%	10.96%	690.571	344.971

Table 8 Conversion Rates & Page Views of Different Restaurant Types

Part 3: Movie Rating Predictions

In the 'CF' tab of the spreadsheet, the sum of squared errors (for the whole training set) is 180 and this is the benchmark we would like to beat.

Collaborative Filtering Approach

Recall that the Pearson Correlation is defined as the below formula:

$$w_{a,u} = \frac{\sum_{i \epsilon I_a \cap I_u} (r_{a,i} - \bar{r}_a)(r_{u,i} - \bar{r}_u)}{\sqrt{\sum_{i \epsilon I_a \cap I_u} (r_{a,i} - \bar{r}_a)^2} \sqrt{\sum_{i \epsilon I_a \cap I_u} (r_{u,i} - \bar{r}_u)^2}}$$

In the above formula, the average user ratings $\bar{r}_a$ and $\bar{r}_u$ are defined as the average ratings for user a and user u. All the available ratings for a given users are used for the calculation. For example, $\bar{r}_a$ is calculated by averaging all the movie ratings given by user a, regardless of whether the respective movies are also rated by user u.

In the PERSON formula of Excel, only the movie ratings that are rated by both user a and user u are used in the calculation of $\bar{r}_a$ and $\bar{r}_u$. Hence, we shall recalculate the PEARSON correlation using the correct formula to see if it yields better prediction results. Cosine similarity has also been tested and results are tabulated in the below table.

Variation to the Collaborative Filtering	In-Sample Sum of Squared Errors
Cosine Similarity	345.20
Pearson Correlation (corrected $\bar{r}_a$ and $\bar{r}_u$)	177.61

Table 9 Performance of Different Similarity Measures

From the above table, it can be seen that cosine similarity produces significantly worse results while the corrected Pearson produces slightly better sum of squared errors (compared to the benchmark), around 1.5% improvement. Therefore, it is evident that Pearson is able to better capture similarity of user tastes (than cosine similarity) and the corrected Pearson produces the better estimation.

In these approaches, all the other users' ratings on movie i are used in estimating the rating of user a on movie i, $p_{a,i}$. We can also make a slight tweak such that only user with strong similarities are used in the final estimation. The results are tabulated below:

Variation to the Collaborative Filtering	In-Sample Sum of Squared Errors
- **Pearson Correlation (corrected $\bar{r}_a$ and $\bar{r}_u$)** - **Only use the neighbours u of a if $w_{a,u} \geq 0$**	240.32
- **Pearson Correlation (corrected $\bar{r}_a$ and $\bar{r}_u$)** - **Only use the neighbours u of a if $\lvert w_{a,u} \rvert \geq 0.25$**	167.61
- **Pearson Correlation (corrected $\bar{r}_a$ and $\bar{r}_u$)** - **Only use the 6 nearest neighbours of a with highest $\lvert w_{a,u} \rvert$**	166.44

Table 10 Performance of Different Neighbour Selection Methods

In the first scenario of Table 10, we only use the neighbours who are positively correlated with user a. In other words, we are discarding those users with opposite taste as user a in the prediction model. This approach yields worse results than the benchmark. It can be deduced that the users with opposite tastes actually increase the predictive power of the model.

Therefore, in the second scenario, we select only the neighbours u who have $|w_{a,u}| \geq 0.25$. In other words, we discard those users who are weakly positively correlated or weakly negatively correlated with user a. It is expected that the users who are strongly correlated (regardless whether positively or negatively correlated) with user a will produce better predictions. Indeed, the improvement is close to 7%.

Next, in the third scenario, we select only 6 neighbours with the highest absolute correlation, $|w_{a,u}|$. Similar to the earlier argument, users who are have very similar taste or opposite taste are highly predictive and hence shall be used. Using this approach, the improvement is around 7.5%.

Latent Factor Model Approach

The Latent Factor Model (LFM) approach in the Excel spreadsheet uses the "unrefined" model where:

$$r_{ui} = \sum_{j=1}^{k} p_{uj} q_{ji}$$

In this exercise, we would use the better "refined" model where:

$$r_{ui} = \mu + b_i + b_u + \sum_{j=1}^{k} p_{uj} q_{ji}$$

μ: Average of all ratings
b_u: User bias
b_i: Item bias

The below tables tabulate the Sum of Squared Errors (SSE) performance of unrefined and refined LFM model, using Alternating Least Squares (ALS) method. This is a reasonably small matrix and ALS converges pretty quickly (we use 100 iterations). The initial values of P and Q matrices are populated using random values between 1 and 5. Since our goal is to minimise the in-sample SSE, we do not need the regularisation term in the ALS objective function.

The in-sample SSE performance of the unrefined and refined LFM model using ALS method is tabulated below in Table 11.

Number of Latent Factors	Unrefined LFM In-Sample SSE	Refined LFM In-Sample SSE
3	510.84	192.55
4	421.34	137.40
5	331.31	98.29

Table 11 Performance of Latent Factors Model with respect to Number of Latent Factors

Firstly, we can observe that the refined model achieves much higher in-sample SSE than unrefined model in all cases. We can also observe that the number of latent factors increases, the SSE decreases. This is expected since we are using in-sample SSE as matric. As we include more latent factors, the model uses more variables to fit the training data and in-sample SSE decreases. If we increase the number of latent factors to 20, the in-sample SSE is essentially zero. This is the extreme case of overfitting and will likely not perform well using another test dataset. To objectively assess the model's ability in generalising to new movie ratings data, a hold-out sample will need to be used (out of scope for this assignment).

From the above table, it can be seen that the 4-factor and 5-factor "refined" LFM models performs reasonably well (in terms of in-sample SSE) while the number of factors used is still reasonably small (hence reducing the risk of overfitting). The respective improvements over the benchmark are 24% and 45% respectively.