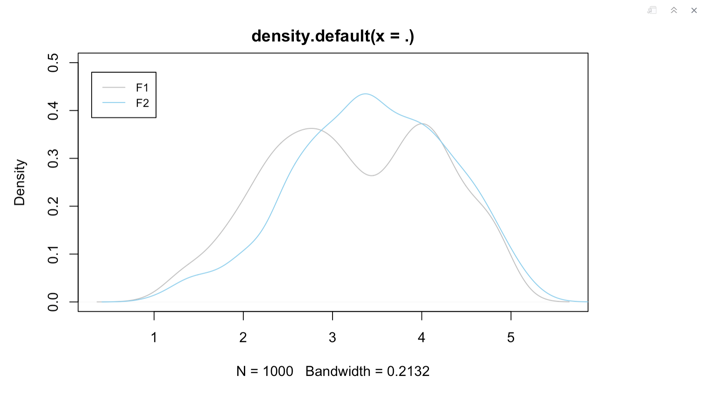
Design Process Report

# Task Abstraction and Domain Situation

|  |  |
| --- | --- |
| Domain Situation | A Marketing analysis for certain company whose |
| Data Tasks Abstraction | **Data:** supplemented include both demographic information and their attitude towards smart watch. The latter is measured by letting them rate importance of a smartwatch feature in a scale of seven.  **Type of abstraction**: explanatory  **Key message**:  “*Of the identified costumer’s cohort, one is make-up of younger cohort, who are also more likely to be Amazon Prime Owners. They are looking for both F1 and F2. It is not clear about the other. Survey question used may not accurately describe what these customer wants.”* |
| Visual Encoding | Kernel Density Plot. |
| Algorithm | Python seaborn library (final version) |

## Statistic:

To briefly summaries the analytic process, first using factor analysis, a dimension reduction technique which have reduced the dimension from seven to two (F1, and F2). Then, ANOVA (Analysis of Variance) and linear regression were used to test if there were any links between demographic group and each factor. In addition, the analysis has done chi-square test to test independence between each categorical variable.



Distribution of F1 is abnormal in a twin bell shape

Several statically problem were encountered. These includes one of the factor score extracted were abnormal and some of ANOVA assumption were not satisfied. The worst of them all is the amount of information that were required to report. Due to statistical analytic reasons, total 18 pair of analysis were done. In the end, the key the analysis was able to distinguish customer preference for one group, but uncertain about the other.

# Visualisation Design Process:

## Evaluating alternatives

### Box Plot or violine plot

Diagram

Description automatically generatedA picture containing text, tool

Description automatically generated

It has been recommended that violine plot is very good to show underlaying distribution. Statistically this is important because of equal variance and normal error assumption in most statistical models. Violine plot is successful showing this. In this case, category “AmznP” and “Degree” in “F1” (first and third plot in first row), were testes to be not satisfying the assumptions. Such violation of assumptions is only illustrated in violin plot where both plots were shown to be in a shape of a “violine” rather than normal “leaf”.

For scientific rigorous, it is not recommended to adapt mathematic modelling if certain assumption were violated but in marketing analysis, one is interested in effect size of each categorical variable. Also, ANOVA is robust according to statisticians (Carmone,1978). Because one is interested in effect size, violine plot may not be suitable. For instance, “AmaznP” in “F2” (first plot of the second row), although weak, is tested statistically significant. It is difficult to tell from violine plot. It is much easier to spot the difference of mean in in Boxplot.

The fact that Audience for Marketing analysis is more interested in effect size than statistic assumptions limited the amount of “key information” one wish to communicate. Violine plot, although appears to be more sophisticated, may cause unnecessary cognitive load. Hence, box plot is better design choice in this domain situation.

### Crossed Box Plot or Interaction Plot

One may wish to find out if combined effect exists since data cross. For instant, would female amazon prime owners factor score higher than male non-amazon prime owner. There are two ways to visualise this. A typical way is to use wrapped boxplot. This is not as effective as using interactive plot, which is more commonly appeared in literatures (all 12 pair of interaction were shown on left)

Chart, box and whisker chart

Description automatically generatedDiagram, engineering drawing

Description automatically generated

Again, the interaction plot reduces cognitive loads by removing visual embodiment. Further, users will have to move eye from one box to another if they wish to compare between different categories. Further, the facet plots the left only show two pairs of interactions.

Interaction plot also has an advantage at algorithm level. To achieve box plot effect on the left requires a lot of coding. The facet grid option from “plot” allows plot each data at different dimension but fails to print two different plot side by side. In the end I must load library “grid extra” to sew two facet plots tougher. In the end it was a lot of code lines. Interaction plot, on the other hand, just requires online for each plot.

### Scatter Plot or Conceptual Plot

One demographic attribute, age, is a continuous variable. As a standard for plotting relationships, scatter plot with hues were used initially.

Scatter chart

Description automatically generated

It is easy to observe that relationship between age and F1 is overall stronger than F2. Further the effect is even more signified when respondent is an amazon prime owner (AmznP = 1). For F2, cross effect is weak. The regression line draw appears almost plateau. This suggest there is little relations between those variables. In marketing analysis, this indicates **survey question used to derive F2 may not accurately describe what these customer wants**.

However, without any verbal aids and careful consideration, it is difficult to draw these conclusions. User may struggle to tell which two plot to compare. If they do compare, their eyes must constantly move up and down to compare one and another. This is the very same problem with interaction plot in the last section. There were already 12 interaction plots. Psychologist suggest human can only remember maximum 5 unrelated item at a time ().

To this point, the purpose of viz idiom has changed from exploratory to explanatory. User may not have the patience to explore every detailed of the plot. A detailed static description is enough to serve as an evidence to reach conclusion. User’s main concern, as a company, is who are those customers. The message we want to communicate is as following:

“*Of the identified costumer’s cohort, one is make-up of younger cohort, who are also more likely to be Amazon Prime Owners. They are looking for both F1 and F2. It is not clear about the other. Survey question used may not accurately describe what these customer wants.”*

Initially, conceptual visualisation has been considered to visualise such idea. But conceptual visualisation would have been less convincing since it does not derive from actual data. It will most certainly defy modern management ideology, “evidence-based decision making” ().

### Kernel Density Plot

During the analysis phases, it has found that many of categorical variables are inter-related. For instance, those with higher degree are more likely to an amazon prime account; female is more likely to be in lower income brand than male.

Diagram

Description automatically generatedDiagram

Description automatically generatedDiagram

Description automatically generated

# Diagram Description automatically generatedDiagram Description automatically generatedDiagram Description automatically generated with medium confidence

It would have been impossible to explain which categorical variable is the triggering reason. The Kernel plot provides an intuitive answer: pattens became most clear when distinguish by *AmznP* (if customer have amazon account or not), and by age. Zooming into “income kernel density plot (top middle)”, it is easy to see the density plot is still have two peaks. Moving to next plot on left (red plot on top left), those two peaks are separated. If looking at different age (either bottom right or bottom middle), some of the line became almost normal. It became clear that those are younger and older cohort has different preference, indicating lighter colour on one side, and darker colour on the other. In the bottom three plot, different colour schemes are explored. The bottom right viz was selected for publishing.

Diagram

Description automatically generated

Users can clearly see the colour segmentation from without having to eye moving from one plot to another. In terms of Task abstraction, the kernel efficiently commutes the key marketing insight from the analysis:

“*Of the identified costumer’s cohort, one is make-up of younger cohort, who are also more likely to be Amazon Prime Owners. They are looking for both F1 and F2. It is not clear about the other. Survey question used may not accurately describe what these customer wants.”*

The use of colour is carefully considered. In this viz idiom, the deeper the colour, the younger the respondent. This may betray common conception that the younger generation are usually represented by lighter colour, but is essential here because when colour is reversed, it’s less clear to see the segregation between two age group (as shown in bottom mid). Schloss et al (2018) study colour finds “dark is more bias” in terms of reading. The message is the younger cohort had stronger preference; therefore, they should keep deeper colour. Finally, there is a clear advantage in Algorithm. It uses Python instead of R, which Holm et al. (2020) considers as more advance in terms of energy and CPU power.

# Reference

Carmone, F. J., Green, P. E., & Jain, A. K. (1978). Robustness of conjoint analysis: Some Monte Carlo results. Journal of Marketing Research, 15(2), 300-303.

Crawley, M. J. (2012). The R book. John Wiley & Sons.

Holm, H.H., Brodtkorb, A.R. and Sætra, M.L., 2020. GPU computing with Python: Performance, energy efficiency and usability. Computation, 8(1), p.4

Munzner, T. (2014). Visualization analysis and design. CRC press.

Schloss, K. B., Gramazio, C. C., Silverman, A. T., Parker, M. L., & Wang, A. S. (2018). Mapping color to meaning in colormap data visualizations. IEEE transactions on visualization and computer graphics, 25(1), 810-819.