Teesside University

School of Computing, Engineering & Digital Technologies
Middlesbrough TS1 3BA

Module Title: Advance Practice
Module Code: CIS4006-N
Project Topic: A Reflective Report

Name: Chinedu Emmanuel Agwunobi

Student ID: B1555134

1 Table of Contents

2 Introduction

Throughout the immersive three-month internship at Think Pacific, this reflective report encapsulates a transformative journey encompassing my individual experiences, insights gained, and contributions made within the dynamic landscape of data science. Beyond a mere recounting of activities, this narrative embarks on a critical self-assessment, highlighting the profound impact of these experiences on both my academic progression and professional evolution. These formative months have not only enriched my understanding of data processing and machine learning but have also significantly shaped my trajectory, fostering substantial academic growth while laying the groundwork for enhanced professional capabilities. Through an introspective lens, this report delves into the interplay between academic knowledge and real-world application, shedding light on the nuanced lessons learned, pivotal skill developments, and the ethical considerations inherent in this evolving field.

3 Activities Undertaken

Embarking on a transformative three-month internship with Think Pacific, this reflective report encapsulates a journey rich in individual experiences, insights, and contributions within the dynamic realm of data science. Beyond detailing activities, this narrative critically assesses the profound impact of experiences on academic and professional growth.

3.1 About Think Pacific

Think Pacific's dedication to community-driven endeavors in Fiji and Bali was palpable and resonated deeply throughout the placement program. Their overarching goal was to synergize academic objectives, cultural immersion, and collaborative engagements within a vibrant and dynamic environment. This convergence served as the cornerstone for an enriching and holistic experience, fostering growth on multiple fronts.

3.2 Phases of the Internship Program

The internship program was meticulously structured into five distinct phases, each serving as a vital milestone in the learning journey. Under the guidance of mentor Grace Chang, these phases provided a structured pathway for skill enhancement and project execution. The one-on-one meetings served as pivotal checkpoints, facilitating the refinement of strategies and the establishment of developmental goals. Spanning over 12 weeks, these phases encompassed:

3.2.1 Discovery Phase: Immersion into Fiji's Cultural Landscape and Organizational Objectives

The initial three weeks of the internship were dedicated to a deep dive into Fiji's rich cultural tapestry, an essential foundation for effective integration and communication. Delving into the multifaceted aspects of Fiji's culture was a revelatory experience. Fijian traditions, customs, and heritage underscored the communal values prevalent in this vibrant society. Celebrating harmony, respect for elders, and fostering strong family ties are intrinsic to Fijian culture. The warm and welcoming nature, encapsulated in the renowned "bula spirit," permeated every interaction (Wikipedia, 2022).

My exploration included embracing traditional ceremonies like Kava and Meke dances, witnessing firsthand the cultural identity embedded in these rituals. Additionally, art, music, and storytelling unveiled a legacy passed down through generations, shedding light on the profound connection to nature and the spiritual depth ingrained in Fijian life. These events were hosted virtually by Think Pacific.

During this phase, a pivotal group mentor meeting led by Grace Chang set the tone for the internship. This session illuminated the various phases, activities, and expectations, offering invaluable insights to maximize the experience. Furthermore, the weekly self-assessment provided by Think Pacific became a crucial tool for evaluating progress, identifying areas for growth, and setting ambitious new goals.

3.2.2 Discussion Phase: Active engagement in discussions with NGO staff and peers, augmenting my understanding.

During the discussion phase, my involvement was marked by active participation and engagement. This period spanned from the fourth to the sixth week, bustling with a myriad of events and engagements, both internal and external. Networking sessions offered opportunities to establish meaningful connections with colleagues and mentors, nurturing an environment conducive to collaboration and the exchange of knowledge.

A significant focus during this phase was on fostering collaboration. I actively contributed to group discussions, leveraging my skills and expertise to align with collective objectives and drive outcomes. Moreover, sessions were conducted to enhance soft skills, covering diverse topics such as emotional intelligence, public speaking, effective communication, health awareness, self-development, and optimizing one's LinkedIn profile. These engagements enriched not just my professional acumen but also contributed to holistic personal development.

3.2.3 Decision Phase: Selecting and initiating a data science project centered on customer behavior analysis.

During the decision phase, spanning from the seventh to the eighth week, the focus pivoted toward selecting a data science project with a core focus on analyzing customer behavior. This phase was pivotal, involving two one-on-one mentor meetings where I presented and discussed potential action projects. Each participant was encouraged to select two projects and finalize one to pursue in collaboration with their mentor, elucidating motivations and reasons behind the choices made.

My chosen action project, titled "51:47 - Designing a cluster profiling software to analyze customer behavior," was a significant decision. The project's emphasis on utilizing R as the programming language was a deliberate choice aimed at enhancing my expertise in R and utilizing R Studio. Notably, this project marked my initial foray into working with unsupervised machine learning algorithms, signifying an opportunity for substantial skill growth and application.

Simultaneously, this phase marked the onset of progress reviews, with valuable feedback from my academic supervisor, Paul Abley. The insights gained from this initial review highlighted commendable achievements in networking, active collaboration, and an authentic embrace of Fiji's diverse culture. It also emphasized the need to strengthen time management skills and deepen technical proficiency, particularly in leveraging R for the data science project. This feedback became the foundation for refining my approach and setting clearer, more focused goals for the subsequent stages of the internship.

3.2.4 Design Phase: Crafting project elements involving R programming, data analysis, and software development.

During the design phase spanning from the ninth to the eleventh week, the focus was on the implementation and refinement of my action project. The primary project undertaken during this phase was the "Customer Behavior Segmentation" initiative, aimed at exploring and categorizing customer behavior utilizing K-Means, DBSCAN, and Hierarchical Clustering in R-Studio with the Mall Customer Dataset.

The project entailed various critical steps, starting with data preprocessing to scale or normalize the dataset, encompassing customer age, annual income, and spending score. The application of DBSCAN, Hierarchical Clustering, and K-Means algorithms facilitated the segmentation of customers into distinct groups based on spending patterns and demographic attributes.

Evaluation metrics, including scatter plots, pair plots, and cluster comparison charts, played a crucial role in assessing the quality and effectiveness of the generated clusters. The project aimed to compare the performance of these models, highlighting their strengths and weaknesses in handling different cluster shapes and sizes, noise, and outliers within the dataset.

The overarching goal was to provide comprehensive insights into customer segmentation methodologies, aiding in the identification of the most suitable model for extracting meaningful and actionable behavior patterns from the Mall Customer Dataset.

Moreover, during this phase, I engaged in my third mentor meeting, a group session where mentors assessed the progress of our action projects and addressed any outstanding queries or challenges, fostering an environment of collaborative learning and support among peers.

3.2.5 Delivery Phase: Presenting the project to Think Pacific's senior management and our partner organization.

During the delivery phase, which culminated in the 12th week, I engaged in several critical activities to ensure the successful completion and presentation of my action project. This phase commenced with a pivotal one-on-one mentor meeting, focusing on project overview, submission procedures, and necessary documentation for the culmination of the internship.

Part of this final phase involved the presentation of our action projects via video format, providing an opportunity to showcase our comprehensive understanding and execution of the project objectives. Files submitted were; R-script file, R-studio notebook file (html file), csv file (Mall dataset), and mp4 file (video presentation).

Additionally, this phase included my second and final progress review with my academic supervisor. This review was instrumental in highlighting the progress made throughout the internship, particularly emphasizing the successful implementation of clustering models using R, which received project approval, signifying a significant milestone in its progression.

As part of the closing activities, we engaged in enriching cultural experiences, marking the culmination of the internship program. The conclusion involved the completion and submission of my action project, ensuring all required files were provided. Furthermore, feedback was actively provided to aid Think Pacific in enhancing future iterations of the internship program.

4 Contribution to the Organisation

- Facilitating Networking: I was actively involved fostering an environment conducive to collaboration and knowledge exchange by engaging in active networking endeavors.
- Data Science Project Engagement: Deeply engaging in a data science project that harmonized academic coursework with organizational objectives, exhibiting a dedication to technical proficiency and practical application.
- Cultural Integration: Embracing Fiji's rich cultural nuances, actively contributing to effective workplace communication and seamless integration within a diverse cultural setting.

5 Reflection

The internship served as a catalyst for substantial personal and professional growth:

- Networking and Collaboration: Expanding my ability to network and collaborate within diverse teams, cultivating a more profound understanding of teamwork dynamics and collaborative work environments.
- Technical Expertise: Significantly enhancing technical proficiency, particularly in harnessing R for data science projects, thereby strengthening my analytical and problem-solving skill set.
- Cultural Sensitivity: Gaining a heightened sense of cultural sensitivity and adaptability, fostering effective communication and cultural integration in a cross-cultural setting.

6 Personal Benefits

This internship became a pivotal avenue for fostering my personal development:

- Continuous Self-assessment: Engaging in regular reflections enabled me to continuously assess and refine my strengths while identifying areas for ongoing improvement.
- Time Management and Cultural Awareness: The experience not only honed my time management skills but also deepened my appreciation for cultural diversity. These competencies are vital for personal growth and effective collaboration.
- Technical Proficiency: The hands-on utilization of tools like R, R Studio, and Unsupervised Machine Learning algorithms formed a substantial part of my learning curve, significantly enhancing my technical skill set.

7 Professional Values and Behaviour

My approach throughout the internship echoed steadfast professional values:

- Consistent Professionalism: I ensured a consistently professional demeanor in every interaction, maintaining high standards within the organization and during collaborative initiatives.
- Engaged Participation: Actively participating in group discussions, meetings, and mentor sessions showcased my dedication and passion, amplifying my contributions throughout the internship journey.

8 Challenges and Future Work

In reflecting on the program, several challenges emerged, providing valuable learning opportunities. Implementing Deep Learning (DL) using R posed a significant challenge due to the language's conventional use for statistical analysis rather than deep neural network implementation. This discrepancy highlighted a gap in readily available resources and libraries for DL within R, making it a complex undertaking.

The exploration of DL within R stands as a future work in progress, presenting an opportunity for further research and skill development. Future endeavors will involve delving deeper into alternative methodologies or considering cross-platform integration to facilitate DL in R effectively. This challenge underscores the evolving nature of data science tools and methodologies, prompting an ongoing commitment to explore and adapt to emerging technologies.

9 Conclusion

The Think Pacific placement program has been transformative, fostering growth in time management, technical proficiency, cultural sensitivity, self-development, and professional etiquette. It provided a valuable opportunity to learn R, enhancing my data science skills and proficiency in software development and analysis. Embracing diverse cultures and seeking continuous self-improvement were key aspects, fostering deeper relationships and a more inclusive work environment.

This experience has been instrumental in refining my approach to data science, teamwork, and adaptability. Facing challenges during the program has significantly enhanced my skill set, laying a strong foundation for future academic pursuits and professional endeavors. The multifaceted learning experiences amalgamating theoretical knowledge with practical applications have been invaluable, enriching my journey in the realm of Data Science.

10 References

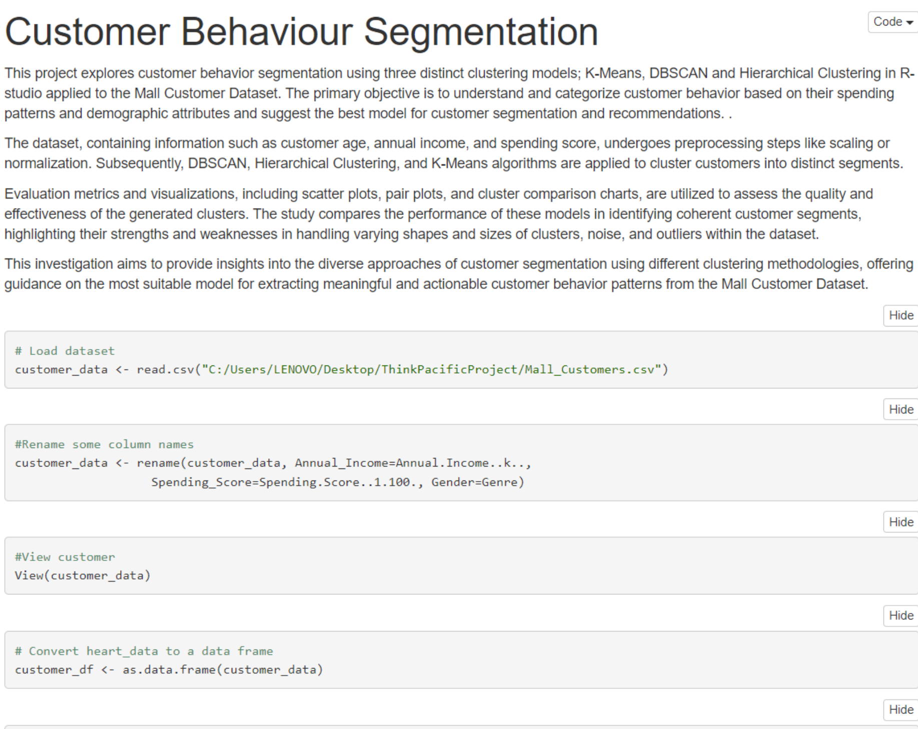
The Library, (2024). Reflective Writing. https://libguides.tees.ac.uk/reflective

Learning Hub, (2024). Reflective Writing
https://www.tees.ac.uk/depts/lis/learninghub/reflective_writing/story_html5.html

Wikepedia, (2022). Culture of Fiji. https://en.wikipedia.org/wiki/Culture_of_Fiji

11 Appendices

Some of the snippets from the Project.



Customer Behaviour Segmentation

Code

This project explores customer behavior segmentation using three distinct clustering models; K-Means, DBSCAN and Hierarchical Clustering in R-studio applied to the Mall Customer Dataset. The primary objective is to understand and categorize customer behavior based on their spending patterns and demographic attributes and suggest the best model for customer segmentation and recommendations. .

The dataset, containing information such as customer age, annual income, and spending score, undergoes preprocessing steps like scaling or normalization. Subsequently, DBSCAN, Hierarchical Clustering, and K-Means algorithms are applied to cluster customers into distinct segments.

Evaluation metrics and visualizations, including scatter plots, pair plots, and cluster comparison charts, are utilized to assess the quality and effectiveness of the generated clusters. The study compares the performance of these models in identifying coherent customer segments, highlighting their strengths and weaknesses in handling varying shapes and sizes of clusters, noise, and outliers within the dataset.

This investigation aims to provide insights into the diverse approaches of customer segmentation using different clustering methodologies, offering guidance on the most suitable model for extracting meaningful and actionable customer behavior patterns from the Mall Customer Dataset.

Hide

```
# Load dataset
customer_data <- read.csv("C:/Users/LENOVO/Desktop/ThinkPacificProject/Mall_Customers.csv")
```

Hide

```
#Rename some column names
customer_data <- rename(customer_data, Annual_Income=Annual.Income..k..,
                   Spending_Score=Spending.Score..1.100., Gender=Genre)
```

Hide

```
#View customer
View(customer_data)
```

Hide

```
# Convert heart_data to a data frame
customer_df <- as.data.frame(customer_data)
```

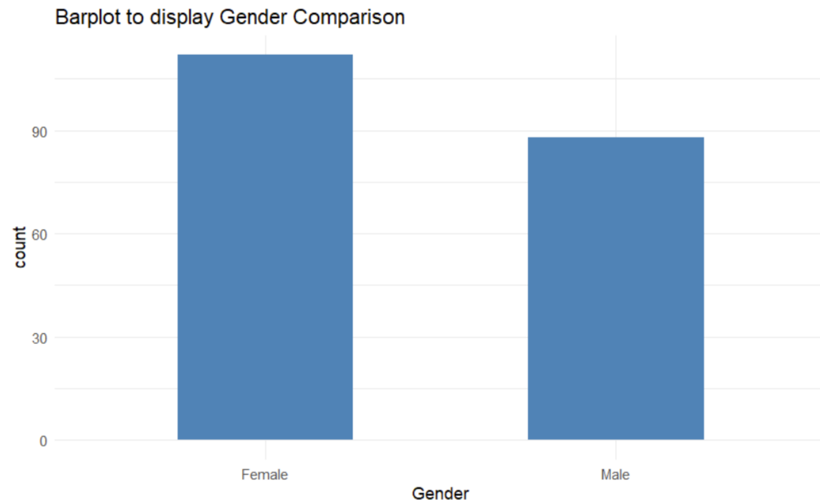
Hide

Fig. 1. Loading of the data in R- Studio Notebook

Exploratory Data Analysis(EDA)

Hide

```r
# Creating a barplot to assess gender distribution of my sample of customers.
ggplot(customer_df,aes(x= Gender)) +
  geom_bar(stat="count",width=0.5,fill="steelblue") +
  theme_minimal()+
  labs(title="Barplot to display Gender Comparison", xlab="Gender")
```



F From the bar chart, we can see that females are in the lead with a share around 56% whereas the males have a share around 44%.

Fig. 2. Bar Plot showing the count of Male and Female Customers

```r
# Calculate the correlation matrix
numeric_customer_df <- customer_df[, sapply(customer_df, is.numeric)]
corr_matrix <- cor(numeric_customer_df)

# Heatmap
ggplot(data = melt(corr_matrix), aes(Var1, Var2, fill = value)) +
  geom_tile() +
  geom_text(aes(label = round(value, 2)), vjust = 1) +  # Add labels with 2 decimal places
  scale_fill_gradient(low = "steelblue", high = "gray") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Correlation Heatmap")
```

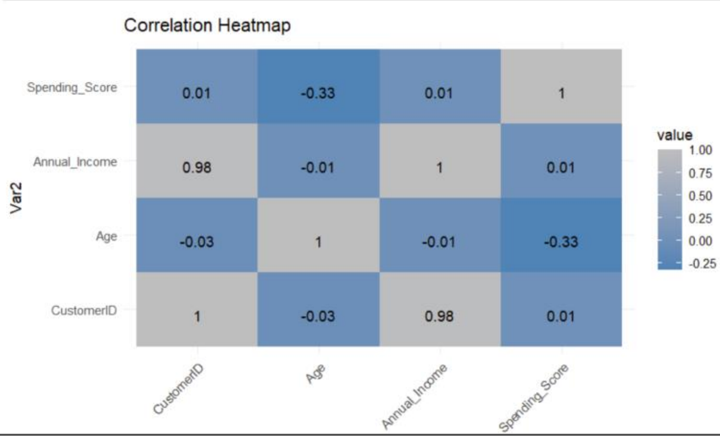


Fig. 3. A Correlation Heat map for the Numerical Variables

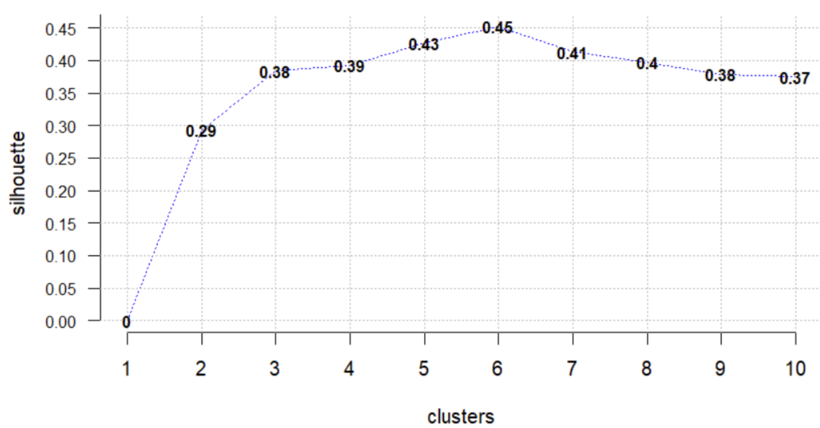
```
# Assuming 'customer_df' is your data frame
ggplot(customer_df, aes(x = Spending_Score, y = Gender, fill = Gender)) +
  geom_boxplot() +
  labs(title = "Boxplot showing customers' Spending Score by Gender") +
  scale_fill_manual(values = c("orange","steelblue")) +
  theme_minimal()
```

From the boxplot, we can see that the median spending score for both males and females are equal. We can also see that more women have a spending score above the median (50), whereas men tend to have a spending score below the median.

Fig. 4 A boxplot showing the Spending Score for both Male and Female Customers

Hide

```
# Getting the optimal clusters
opt <- Optimal_Clusters_KMeans(customer_df[, 3:5], max_clusters = 10, plot_clusters = T, criterion = 'silhouette')
```



The highest average sillhoute value (equal to 0.45) is present for k = 6. Therefore we should opt for 6 clusters in our further analysis with k-means algorithm.

Fig. 5. Graph to get the number of Clusters needed to train the K-Means Model

Creating the k-means clustering model for the data

Hide

```r
#Creating the customer clusters with KMeans
km6<-kmeans(customer_df[,3:5], 6, iter.max = 100, nstart=50,
          algorithm = "Lloyd")
#Print the result
km6
```

```
K-means clustering with 6 clusters of sizes 45, 35, 38, 21, 39, 22

Cluster means:
       Age Annual_Income Spending_Score
1 56.15556      53.37778       49.08889
2 41.68571      88.22857       17.28571
3 27.00000      56.65789       49.13158
4 44.14286      25.14286       19.52381
5 32.69231      86.53846       82.12821
6 25.27273      25.72727       79.36364

Clustering vector:
  [1] 4 6 4 6 4 6 4 6 4 6 4 6 4 6 4 6 4 6 4 6 4 6 4 6 4 6 4 6 4 6 4 6 4 6 4 6 4 6 4 6 1 6 1 3 4 6
 [47] 1 3 3 3 1 3 3 1 1 1 1 1 3 1 1 3 1 1 1 3 1 1 3 3 1 1 1 1 1 3 1 3 3 1 1 3 1 1 3 1 1 3 3 1 1 3
 [93] 1 3 3 3 1 3 1 3 3 1 1 3 1 3 1 1 1 1 1 3 3 3 3 3 1 1 1 1 3 3 3 5 3 5 2 5 2 5 2 5 3 5 2 5 2 5
[139] 2 5 2 5 3 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5
[185] 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5

Within cluster sum of squares by cluster:
[1]  8062.133 16690.857  7742.895  7732.381 13972.359  4099.818
 (between_SS / total_SS =  81.1 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"
[7] "size"         "iter"         "ifault"
```

Fig. 6. Summary of the K-Means Model

```r
ggplot(customer_df, aes(x = Annual_Income , y = Spending_Score)) +
  geom_point(stat = "identity", aes(color = as.factor(km6$cluster))) +
  scale_color_discrete(name = " ",
                       breaks=c( "6", "1", "4", "3", "2","5"),
                       labels=c("Low Income, High Spending", "Medium Income, Medium Spending", "Low Income, Low Spending",
                                "Medium Income, Medium Spending","High Income, Low Spending", "High Income, High Spending"))
+
  labs(x="Annual Income", y="Spending Score") +
  ggtitle("Segments of Mall X Customers",
          subtitle = "Using K-means Clustering")
```

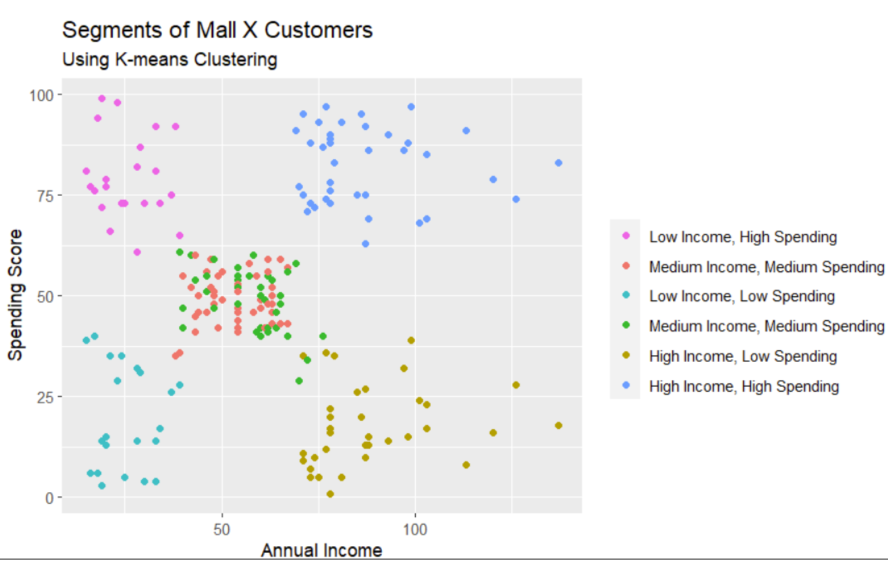


Fig. 7. Descriptive Plot of K-Means Clustering Model

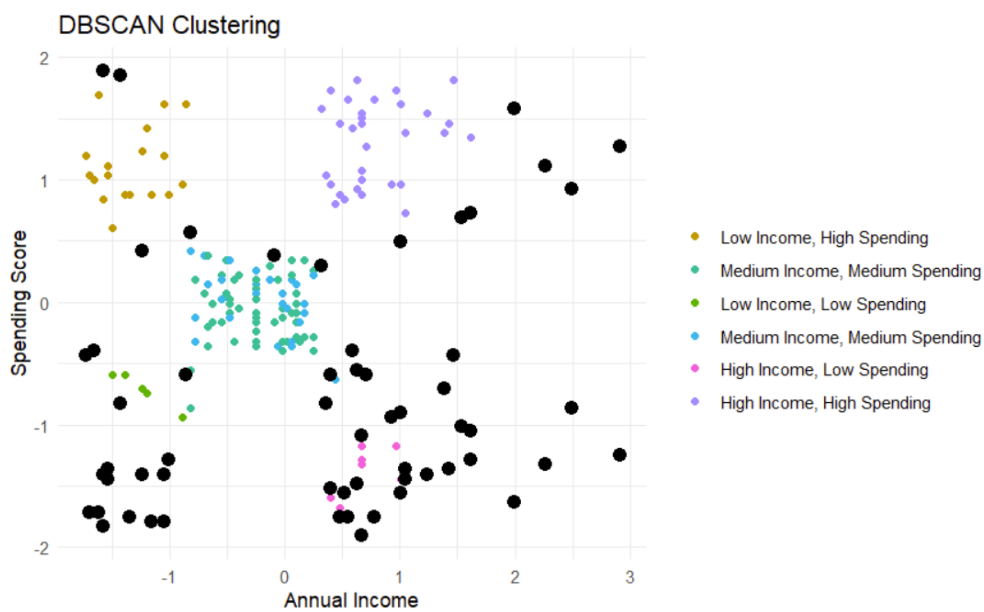


Fig. 8. Descriptive Plot of DBSCAN Clustering Model

Fig. 9. Descriptive Plot of Hierarchical Clustering Model

```r
# Getting the characteristcs each of the groups
segment_customers %>% group_by(cluster, Gender) %>%
  summarise_all(list(mean)) %>% arrange(cluster)
```

cluster <int>	Gender <chr>	CustomerID <dbl>	Age <dbl>	Annual_Income <dbl>	Spending_Score <dbl>
1	Female	24.71429	43.21429	27.35714	21.71429
1	Male	20.33333	48.33333	24.66667	19.66667
2	Female	20.83333	25.58333	24.58333	81.83333
2	Male	23.00000	23.75000	25.50000	79.75000
3	Female	86.72000	40.20000	55.32000	48.92000
3	Male	82.66667	45.12121	53.90909	51.06061
4	Female	163.33333	32.19048	86.04762	81.66667
4	Male	160.44444	33.27778	87.11111	82.66667
5	Female	171.00000	44.60000	92.33333	21.60000
5	Male	159.50000	39.50000	85.15000	14.05000

1-10 of 10 rows

Show

In the first cluster we have middle aged women and men, whose both annual income and spending scores are small. In the second group we have young women and men, who despite the fact they don't have much income, tend to spend a lot. The third group is the most numerous one, which was right in the middle of the presented plots. This cluster constitues of female and male in their forties, which get a middle-sized wages and have moderate spending habits. In group number 4, there are mostly people in their early 30s who earn a lot and also tend to spend much. In the last cluster (number 5), we can see women whose average age was around 44 years old and men with average age circa 39 years old. This group of people, similarly to those in group number 4, have high annual incomes, but on the contrary, they do not like to spend much.

All in all this table shown us, how many important information we can get from the clustering analysis. This paper analysed only a very basic, two dimensional example, which would not be the scenario in most of the business use cases. Nevertheless, I believe that these methods, with some other additional analysis, can be successfully implemented to real business problems for tracking relevant customers.

Fig. 10. Summary of the Hierarchical Clustering Model