UNDERSTANDING SPENDING HABITS OF GEN Z AND MILLENNIALS USING MULTIVARIATE AND PREDICTIVE MODELING TECHNIQUES

Manzano, Wilma Ellaine (301181591)

Rimando, Karon Mae (301168240)

Business Analytics and Insights, Centennial College

BA 723: Business Analytics Capstone

Professor David Parent and Professor Resmi Ann Thomas

August 12, 2022

TABLE OF CONTENTS

EXECUTIVE SUMMARY

This report aims to explore consumer behaviour of Gen Z and millennials and understand its relation to their spending patterns in consideration of several factors including their demographics, preferences, hobbies, interests and personality traits. Through the Exploratory Factor Analysis multivariate technique, latent variables and relevant groupings were identified and used for dimension reduction and model simplification. Subsequently, classification techniques namely Decision Tree and Logistic Regression were employed for identifying the classification between high spenders versus low spenders, and the significant variables that poses a strong impact on spending habits were identified and interpreted accordingly. Certain personality traits were found to have strong influence on spending, such as being trendy, sociable, trustworthy, conscientious along with various interests in art, physical activity, and in computer technology. Males and educational attainment were found to be significant factors too, with males being 28% more likely to become high spenders than females and those who have completed college/ university degree are expected to be more financially savvy. Cluster analysis was also employed to reveal customer segmentation to aid companies customize their marketing campaigns. Results from the segmentation are aligned with the identified classification and likelihood from the Logistic Regression model, where males and those who are highly extroverted, trendy, and physically active were profiled to be the most spenders. Insights derived from both analyses will be beneficial for ecommerce companies on how they can better reach out and target the identified segments.

1. INTRODUCTION

1.1 Background

The United Nations reported there are about 1.2 billion young people today which accounts to roughly 16% of the global population (Youth, 2022). Through this data, we can expect a number of opportunities and potential as the young adults, Gen Z and millennials, tend to have an encouraging mindset towards building a better future. In fact their role today is vital as they are considered to be the main agents of change and progress.

One of the fundamental skills that the young must take into consideration though is having financial literacy as it will benefit them now and in the future in terms of making sound financial decisions. While saving money is an ultimate aspect to financial independence, spending it wisely is of equal importance.

This project aims to better understand the spending patterns of a young person living in Slovakia, a country considered as one of the youthful countries in Europe, based on their the demographics along with their preferences, hobbies, interests and personality traits. In fact, according to Minns, young people in Slovakia represents 34.17% of their total population (Rural Neets in Slovakia, 2022). Given that this age group represents a high proportion and considered as the next in-line in terms of purchasing power, we aim to provide an insight that will help ecommerce businesses, such as H&M, as they continue to explore means that will help them engage and win over the hearts and loyalty of the young consumers through marketing and advertising.

The ecommerce industry significantly accelerated and reached its peak at the time of the pandemic as people reverted to online shopping for safety, ease and convenience. There is no doubt that it will continue to rise over the years for several reasons including: personalization, mobility and convenience, more varieties, and its

integration with social media for a wider reach; and so understanding its growth will enable businesses to penetrate into this emerging segment in a more meaningful way.

In 2021, Slovakia was reported to be the 59th largest market for ecommerce with its revenue amounting to $1.4 billion US dollars (ecommerceDB, 2022). Statistics show that there are currently 3.5 million ecommerce users in the country and is expected to grow as new markets and new trends continue to emerge. According to the same report published by ecommerceDB, Fashion seems to be the top segment in Slovakia at 34%, followed by Electronics and Media at 29%, Toys/Hobby/DIY at 18%, Food & Personal Care at 10%, and the remaining percentage accounts for Furniture and Appliances (ecommerceDB, 2022).

Although the presence of the global ecommerce website, H&M, is available in Slovakia, local eCommerce websites remain to be popular such that the top three stores: alza.sk, mall.sk and itesco.sk contributed to 50% of online revenue in Slovakia in 2021. While the ecommerce business in the country remains to be favourable due to its large and active consumer market, H&M should explore its full potential by taking into consideration the components of localization where it will lead them to opportunities such as increase in customer engagement and conversion. One way is through understanding Gen Z and millennials' spending habits better and looking into the best segments they could explore and potentially target.

1.2 Problem Statement

This study aims to identify factors from demographics, preferences, interests, and personality traits that would best explain Gen Z and millennials' spending patterns. The insight that will be derived from this study will be beneficial for H&M as it will help them optimize their customer service to the fullest, allowing them to enhance their offerings and giving them opportunity to improve their strategies in keeping up with the

latest trends through marketing, and more importantly boosting their sales while increasing their online presence and reach in this growing segment through advertising.

Maximizing data to its full potential especially in the ecommerce world is vital because it empowers businesses as they gain relevant insights into customer behaviour and at the same time opens up a number of opportunities for growth. It would be ideal for businesses to utilize data to its highest capacity so as to prevent any missed opportunities leading towards progression.

1.3 Objectives and Measurement

The main analytics objective of this study is exploring which among demographic profiles, preferences, hobbies, interests, and personality traits are good predictors of a young adult's spending habits. Specifically, the null and alternative hypothesis are as follows:

Null Hypothesis (*Ho*): There is no significant relationship between a young adult's spending habits and his/ her music preferences, movie preferences, hobbies, interests, and personality traits.

Alternative Hypothesis (*Ha*): There is a statistically significant relationship between a young adult's spending habits and his/ her music preferences, movie preferences, hobbies, interests, and personality traits.

The number of predictors will also be summarized into smaller representative variables for model simplification and ease of interpretation. Respondents would also be segmented to aid the ecommerce company with their customer targeting.

The following metrics will be used to determine the performance of the model:

1. Overall goodness of fit of regression model: *low average squared error, accuracy, f-score and their gap between training and validation dataset*
2. Number of independent variables: *least number of variables that has the most optimal fit is desired*
3. Number of clusters identified: *4 to 5 customer segments that can be easily targeted*

1.4 Assumptions and Limitations

Some assumptions and limitations regarding the data are outlined to be considered for the analysis and interpretation.

1. Survey data was not sampled randomly, rather through convenience-snowball only. Due to this, the sample distribution may not reflect actual population census.
2. Spending patterns are based on Likert scale (1 as strongly disagree, and 5 as strongly agree), instead of actual purchase spending.

2. DATA SOURCE

2.1 Data Introduction

Data used for this analysis was gathered by Statistics students from a university in Bratislava, Slovakia, and retrieved from Kaggle.com. The study was done in 2013 via snowball sampling, wherein the students asked their friends to participate in the survey and complete either via electronic or written forms. All of the participants were of Slovakian nationality, aged between 15 to 30 years old. The original questionnaire was in Slovak language and was later translated into English, covering questions on the following topics:

- *Music preferences*

- *Movie preferences*
- *Hobbies & interests*
- *Phobias*
- *Health habits*
- *Personality traits, views on life, & opinions*
- *Spending habits*
- *Demographics*

2.2 Exclusions

There were no exclusions reported for this data.

2.3 Data Dictionary

Variable Name	Label	Response Options	Level
MUSIC PREFERENCES			
Music	I enjoy listening to music.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Slowsongsorfastsongs	I prefer Slow paced music/ Fast paced music	Slow paced music 1-2-3-4-5 Fast paced music	integer
Dance	Dance, Disco, Funk	Don't enjoy at all 1-2-3-4-5 Enjoy very much	integer
Folk	Folk music	Don't enjoy at all 1-2-3-4-5 Enjoy very much	integer
Country	Country	Don't enjoy at all 1-2-3-4-5 Enjoy very much	integer
Classicalmusic	Classical	Don't enjoy at all 1-2-3-4-5 Enjoy very much	integer
Musical	Musicals	Don't enjoy at all 1-2-3-4-5 Enjoy very much	integer
Pop	Pop	Don't enjoy at all 1-2-3-4-5 Enjoy very much	integer
Rock	Rock	Don't enjoy at all 1-2-3-4-5 Enjoy very much	integer
MetalorHardrock	Metal, Hard rock	Don't enjoy at all 1-2-3-4-5 Enjoy very much	integer
Punk	Punk	Don't enjoy at all 1-2-3-4-5 Enjoy very much	integer
HiphopRap	Hip hop, Rap	Don't enjoy at all 1-2-3-4-5 Enjoy very much	integer

ReggaeSka	Reggae, Ska	Don't enjoy at all 1-2-3-4-5 Enjoy very much	integer
SwingJazz	Swing, Jazz	Don't enjoy at all 1-2-3-4-5 Enjoy very much	integer
Rocknroll	Rock n Roll	Don't enjoy at all 1-2-3-4-5 Enjoy very much	integer
Alternative	Alternative music	Don't enjoy at all 1-2-3-4-5 Enjoy very much	integer
Latino	Latin	Don't enjoy at all 1-2-3-4-5 Enjoy very much	integer
TechnoTrance	Techno, Trance	Don't enjoy at all 1-2-3-4-5 Enjoy very much	integer
Opera	Opera	Don't enjoy at all 1-2-3-4-5 Enjoy very much	integer
MOVIE PREFERENCES			
Movies	I really enjoy watching movies.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Horror	Horror movies	Don't enjoy at all 1-2-3-4-5 Enjoy very much	integer
Thriller	Thriller movies	Don't enjoy at all 1-2-3-4-5 Enjoy very much	integer
Comedy	Comedies	Don't enjoy at all 1-2-3-4-5 Enjoy very much	integer
Romantic	Romantic movies	Don't enjoy at all 1-2-3-4-5 Enjoy very much	integer
Scifi	Sci-fi movies	Don't enjoy at all 1-2-3-4-5 Enjoy very much	integer
War	War movies	Don't enjoy at all 1-2-3-4-5 Enjoy very much	integer
FantasyFairytales	Tales	Don't enjoy at all 1-2-3-4-5 Enjoy very much	integer

Animated	Cartoons	Don't enjoy at all 1-2-3-4-5 Enjoy very much	integer
Documentary	Documentaries	Don't enjoy at all 1-2-3-4-5 Enjoy very much	integer
Western	Western movies	Don't enjoy at all 1-2-3-4-5 Enjoy very much	integer
Action	Action movies	Don't enjoy at all 1-2-3-4-5 Enjoy very much	integer
HOBBIES & INTERESTS			
History	History	Not interested 1-2-3-4-5 Very interested	integer
Psychology	Psychology	Not interested 1-2-3-4-5 Very interested	integer
Politics	Politics	Not interested 1-2-3-4-5 Very interested	integer
Mathematics	Mathematics	Not interested 1-2-3-4-5 Very interested	integer
Physics	Physics	Not interested 1-2-3-4-5 Very interested	integer
Internet	Internet	Not interested 1-2-3-4-5 Very interested	integer
PC	PC Software, Hardware	Not interested 1-2-3-4-5 Very interested	integer
EconomyManagement	Economy, Management	Not interested 1-2-3-4-5 Very interested	integer
Biology	Biology	Not interested 1-2-3-4-5 Very interested	integer
Chemistry	Chemistry	Not interested 1-2-3-4-5 Very interested	integer
Reading	Poetry reading	Not interested 1-2-3-4-5 Very interested	integer

Geography	Geography	Not interested 1-2-3-4-5 Very interested	integer
Foreignlanguages	Foreign languages	Not interested 1-2-3-4-5 Very interested	integer
Medicine	Medicine	Not interested 1-2-3-4-5 Very interested	integer
Law	Law	Not interested 1-2-3-4-5 Very interested	integer
Cars	Cars	Not interested 1-2-3-4-5 Very interested	integer
Artexhibitions	Art	Not interested 1-2-3-4-5 Very interested	integer
Religion	Religion	Not interested 1-2-3-4-5 Very interested	integer
Countrysideoutdoors	Outdoor activities	Not interested 1-2-3-4-5 Very interested	integer
Dancing	Dancing	Not interested 1-2-3-4-5 Very interested	integer
Musicalinstruments	Playing musical instruments	Not interested 1-2-3-4-5 Very interested	integer
Writing	Poetry writing	Not interested 1-2-3-4-5 Very interested	integer
Passivesport	Sport and leisure activities	Not interested 1-2-3-4-5 Very interested	integer
Activesport	Sport at competitive level	Not interested 1-2-3-4-5 Very interested	integer
Gardening	Gardening	Not interested 1-2-3-4-5 Very interested	integer
Celebrities	Celebrity lifestyle	Not interested 1-2-3-4-5 Very interested	integer
Shopping	Shopping	Not interested 1-2-3-4-5 Very interested	integer

Scienceandtechnology	Science and technology	Not interested 1-2-3-4-5 Very interested	integer
Theatre	Theatre	Not interested 1-2-3-4-5 Very interested	integer
Funwithfriends	Socializing	Not interested 1-2-3-4-5 Very interested	integer
Adrenalinesports	Adrenaline sports	Not interested 1-2-3-4-5 Very interested	integer
Pets	Pets	Not interested 1-2-3-4-5 Very interested	integer
PHOBIAS			
Flying	Flying	Not afraid at all 1-2-3-4-5 Very afraid of	integer
Storm	Thunder, lightning	Not afraid at all 1-2-3-4-5 Very afraid of	integer
Darkness	Darkness	Not afraid at all 1-2-3-4-5 Very afraid of	integer
Heights	Heights	Not afraid at all 1-2-3-4-5 Very afraid of	integer
Spiders	Spiders	Not afraid at all 1-2-3-4-5 Very afraid of	integer
Snakes	Snakes	Not afraid at all 1-2-3-4-5 Very afraid of	integer
Rats	Rats, mice	Not afraid at all 1-2-3-4-5 Very afraid of	integer
Ageing	Ageing	Not afraid at all 1-2-3-4-5 Very afraid of	integer
Dangerousdogs	Dangerous dogs	Not afraid at all 1-2-3-4-5 Very afraid of	integer
Fearofpublicspeaking	Public speaking	Not afraid at all 1-2-3-4-5 Very afraid of	integer

HEALTH HABITS			
Smoking	Smoking habits	Never smoked - Tried smoking - Former smoker - Current smoker	categorical
Alcohol	Drinking	Never - Social drinker - Drink a lot	categorical
Healthyeating	I live a very healthy lifestyle.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
PERSONALITY TRAITS, VIEWS ON LIFE & OPINIONS			
Dailyevents	I take notice of what goes on around me.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Prioritisingworkload	I try to do tasks as soon as possible and not leave them until last minute.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Writingnotes	I always make a list so I don't forget anything.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Workaholism	I often study or work even in my spare time.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Thinkingahead	I look at things from all different angles before I go ahead.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Finaljudgement	I believe that bad people will suffer one day and good people will be rewarded.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Reliability	I am reliable at work and always complete all tasks given to me.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Keepingpromises	I always keep my promises.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Lossofinterest	I can fall for someone very quickly and then completely lose interest.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Friendsversusmoney	I would rather have lots of friends than lots of money.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Funniness	I always try to be the funniest one.	Strongly disagree 1-2-3-4-5 Strongly agree	integer

Fake	I can be two faced sometimes.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Criminaldamage	I damaged things in the past when angry.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Decisionmaking	I take my time to make decisions.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Elections	I always try to vote in elections.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Selfcriticism	I often think about and regret the decisions I make.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Judgmentcalls	I can tell if people listen to me or not when I talk to them.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Hypochondria	I am a hypochondriac.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Empathy	I am empathetic person.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Eatingtosurvive	I eat because I have to. I don't enjoy food and eat as fast as I can.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Giving	I try to give as much as I can to other people at Christmas.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Compassiontoanimals	I don't like seeing animals suffering.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Borrowedstuff	I look after things I have borrowed from others.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Loneliness	I feel lonely in life.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Cheatinginschool	I used to cheat at school.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Health	I worry about my health.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Changingthepast	I wish I could change the past because of the things I have done.	Strongly disagree 1-2-3-4-5 Strongly agree	integer

God	I believe in God.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Dreams	I always have good dreams.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Charity	I always give to charity.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Numberoffriends	I have lots of friends.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Punctuality	Timekeeping.	I am often early. - I am always on time. - I am often running late.	categorical
Lying	Do you lie to others?	Never. - Only to avoid hurting someone. - Sometimes. - Everytime it suits me.	categorical
Waiting	I am very patient.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Newenvironment	I can quickly adapt to a new environment.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Moodswings	My moods change quickly.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Appearenceandgestures	I am well-mannered and I look after my appearance.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Socializing	I enjoy meeting new people.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Achievements	I always let other people know about my achievements.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Respondingtoaseriousletter	I think carefully before answering any important letters.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Children	I enjoy children’s' company.	Strongly disagree 1-2-3-4-5 Strongly agree	integer

Assertiveness	I am not afraid to give my opinion if I feel strongly about something.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Gettingangry	I can get angry very easily.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Knowingtherightpeople	I always make sure I connect with the right people.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Publicspeaking	I have to be well prepared before public speaking.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Unpopularity	I will find a fault in myself if people don't like me.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Lifestruggles	I cry when I feel down or things don't go the right way.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Happinessinlife	I am 100% happy with my life.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Energylevels	I am always full of life and energy.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Smallbigdogs	I prefer big dangerous dogs to smaller, calmer dogs.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Personality	I believe all my personality traits are positive.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Findinglostvaluables	If I find something the doesn't belong to me I will hand it in.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Gettingup	I find it very difficult to get up in the morning.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Interestsorhobbies	I have many different hobbies and interests.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Parentsadvice	I always listen to my parents' advice.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Questionnairesorpolls	I enjoy taking part in surveys.	Strongly disagree 1-2-3-4-5 Strongly agree	integer

Internetusage	How much time do you spend online?	No time at all - Less than an hour a day - Few hours a day - Most of the day	categorical
SPENDING HABITS			
Finances	I save all the money I can.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Shoppingcentres	I enjoy going to large shopping centres.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Brandedclothing	I prefer branded clothing to none branded.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Entertainmentspending	I spend a lot of money on partying and socializing.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Spendingonlooks	I spend a lot of money on my appearance.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Spendingongadgets	I spend a lot of money on gadgets.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
Spendingonhealthyeating	I will happily pay more money for good, quality or healthy food.	Strongly disagree 1-2-3-4-5 Strongly agree	integer
DEMOGRAPHICS			
Age	Age		integer
Height	Height		integer
Weight	Weight		integer
Numberofsiblings	How many siblings do you have?		integer
Gender	Gender	Female - Male	categorical
Leftrighthanded	I am left/ right-handed	Left handed - Right handed	categorical
Education	Highest education achieved	Currently a Primary school pupil - Primary school - Secondary school - College/Bachelor degree	categorical
Onlychild	I am the only child	No - Yes	categorical

Villagetown	I spent most of my childhood in a	City - village	categorical
Houseblockofflats	I lived most of my childhood in a	house/bungalow - block of flats	categorical

3. DATA EXPLORATION

3.1 Data Exploration Techniques

Initial data investigations were performed on the data via SPSS to check for patterns and discover anomalies that could impact the analysis. Descriptive statistics, frequency tables, and histogram were generated to study the distribution of each variable as shown in the SPSS syntax under *Appendix A.*

The descriptive statistics output table are provided below. No variables have high missing rate, though some of the scale questions such as interest in music, movies, comedy, writing, gardening, having fun with friends, along with other demographic variables on age, height, weight, and number of siblings were found to have high skewness and kurtosis values, which denotes that these variables deviate from normality.

Descriptive Statistics									
	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Music	1007	1	5	4.73	0.664	-3.108	0.077	10.997	0.154
Slow songs or fast songs	1008	1	5	3.33	0.834	0.157	0.077	0.602	0.154
Dance	1006	1	5	3.11	1.171	-0.046	0.077	-0.803	0.154
Folk	1005	1	5	2.29	1.139	0.695	0.077	-0.216	0.154
Country	1005	1	5	2.12	1.076	0.796	0.077	-0.038	0.154
Classical music	1003	1	5	2.96	1.253	0.107	0.077	-0.969	0.154
Musical	1008	1	5	2.76	1.261	0.22	0.077	-0.928	0.154
Pop	1007	1	5	3.47	1.161	-0.383	0.077	-0.704	0.154
Rock	1004	1	5	3.76	1.185	-0.703	0.077	-0.419	0.154
Metal or Hardrock	1007	1	5	2.36	1.373	0.605	0.077	-0.935	0.154
Punk	1002	1	5	2.46	1.301	0.441	0.077	-0.959	0.154
Hiphop, Rap	1006	1	5	2.91	1.376	0.037	0.077	-1.25	0.154
Reggae, Ska	1003	1	5	2.77	1.214	0.156	0.077	-0.901	0.154
Swing, Jazz	1004	1	5	2.76	1.258	0.146	0.077	-0.998	0.154
Rock n roll	1003	1	5	3.14	1.237	-0.109	0.077	-0.917	0.154
Alternative	1003	1	5	2.83	1.347	0.162	0.077	-1.129	0.154
Latino	1002	1	5	2.84	1.328	0.188	0.077	-1.099	0.154

Techno, Trance	1003	1	5	2.34	1.324	0.57	0.077	-0.906	0.154
Opera	1009	1	5	2.14	1.184	0.842	0.077	-0.23	0.154
Movies	1004	1	5	4.61	0.695	-1.904	0.077	3.581	0.154
Horror	1008	1	5	2.79	1.411	0.169	0.077	-1.259	0.154
Thriller	1009	1	5	3.38	1.198	-0.345	0.077	-0.811	0.154
Comedy	1007	1	5	4.49	0.78	-1.595	0.077	2.276	0.154
Romantic	1007	1	5	3.49	1.209	-0.327	0.077	-0.86	0.154
Sci-fi	1008	1	5	3.11	1.312	-0.054	0.077	-1.107	0.154
War	1008	1	5	3.16	1.348	-0.078	0.077	-1.154	0.154
Fantasy/Fairy tales	1007	1	5	3.75	1.182	-0.552	0.077	-0.704	0.154
Animated	1007	1	5	3.79	1.219	-0.675	0.077	-0.615	0.154
Documentary	1002	1	5	3.64	1.133	-0.486	0.077	-0.584	0.154
Western	1006	1	5	2.13	1.139	0.857	0.077	-0.028	0.154
Action	1008	1	5	3.54	1.236	-0.429	0.077	-0.87	0.154
History	1008	1	5	3.21	1.264	-0.09	0.077	-1.013	0.154
Psychology	1005	1	5	3.14	1.258	-0.079	0.077	-1.038	0.154
Politics	1009	1	5	2.6	1.294	0.37	0.077	-0.94	0.154
Mathematics	1007	1	5	2.33	1.352	0.613	0.077	-0.862	0.154
Physics	1007	1	5	2.06	1.227	0.934	0.077	-0.219	0.154
Internet	1006	1	5	4.18	0.921	-0.923	0.077	0.279	0.154
PC	1004	1	5	3.14	1.322	-0.073	0.077	-1.124	0.154
Economy Management	1005	1	5	2.64	1.347	0.332	0.077	-1.078	0.154
Biology	1004	1	5	2.67	1.384	0.409	0.077	-1.071	0.154

Chemistry	1000	1	5	2.17	1.378	0.957	0.077	-0.407	0.155
Reading	1004	1	5	3.16	1.496	-0.152	0.077	-1.39	0.154
Geography	1001	1	5	3.08	1.282	-0.033	0.077	-1.031	0.154
Foreign languages	1005	1	5	3.78	1.141	-0.642	0.077	-0.435	0.154
Medicine	1005	1	5	2.52	1.38	0.545	0.077	-0.905	0.154
Law	1009	1	5	2.26	1.243	0.726	0.077	-0.504	0.154
Cars	1006	1	5	2.69	1.441	0.293	0.077	-1.26	0.154
Art exhibitions	1004	1	5	2.59	1.322	0.396	0.077	-0.962	0.154
Religion	1007	1	5	2.27	1.32	0.69	0.077	-0.708	0.154
Countryside, outdoors	1003	1	5	3.69	1.196	-0.637	0.077	-0.503	0.154
Dancing	1007	1	5	2.46	1.45	0.545	0.077	-1.074	0.154
Musical instruments	1009	1	5	2.32	1.513	0.698	0.077	-1.037	0.154
Writing	1004	1	5	1.9	1.288	1.214	0.077	0.167	0.154
Passive sport	995	1	5	3.39	1.405	-0.331	0.078	-1.169	0.155
Active sport	1006	1	5	3.29	1.504	-0.291	0.077	-1.341	0.154
Gardening	1003	1	5	1.91	1.175	1.207	0.077	0.518	0.154
Celebrities	1008	1	5	2.36	1.27	0.538	0.077	-0.805	0.154
Shopping	1008	1	5	3.28	1.286	-0.181	0.077	-1.056	0.154
Science and technology	1004	1	5	3.23	1.283	-0.174	0.077	-1.007	0.154
Theatre	1002	1	5	3.02	1.325	0.008	0.077	-1.114	0.154
Fun with friends	1006	2	5	4.56	0.737	-1.655	0.077	2.083	0.154
Adrenaline sports	1007	1	5	2.95	1.421	0.045	0.077	-1.287	0.154
Pets	1006	1	5	3.33	1.545	-0.334	0.077	-1.388	0.154

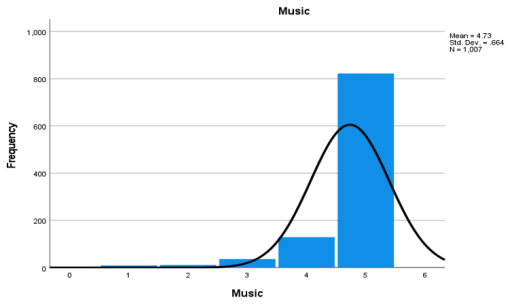
Flying	1007	1	5	2.06	1.211	0.904	0.077	-0.181	0.154
Storm	1009	1	5	1.97	1.164	1.084	0.077	0.262	0.154
Darkness	1008	1	5	2.25	1.255	0.781	0.077	-0.439	0.154
Heights	1007	1	5	2.62	1.295	0.334	0.077	-0.99	0.154
Spiders	1005	1	5	2.83	1.544	0.217	0.077	-1.449	0.154
Snakes	1010	1	5	3.03	1.501	-0.035	0.077	-1.419	0.154
Rats	1007	1	5	2.41	1.401	0.545	0.077	-1.023	0.154
Ageing	1009	1	5	2.58	1.386	0.378	0.077	-1.083	0.154
Dangerous dogs	1009	1	5	3.04	1.367	-0.03	0.077	-1.182	0.154
Fear of public speaking	1009	1	5	2.8	1.215	0.117	0.077	-0.878	0.154
Healthy eating	1007	1	5	3.03	0.937	-0.318	0.077	0.153	0.154
Daily events	1003	1	5	3.07	1.118	0.11	0.077	-0.61	0.154
Prioritising workload	1005	1	5	2.65	1.221	0.26	0.077	-0.821	0.154
Writing notes	1007	1	5	3.08	1.408	-0.063	0.077	-1.267	0.154
Workaholism	1005	1	5	3	1.277	0.059	0.077	-1	0.154
Thinking ahead	1007	1	5	3.41	1.137	-0.19	0.077	-0.808	0.154
Final judgement	1003	1	5	2.65	1.379	0.259	0.077	-1.099	0.154
Reliability	1006	1	5	3.86	0.934	-0.525	0.077	-0.23	0.154
Keeping promises	1009	1	5	3.99	0.899	-0.73	0.077	0.318	0.154
Loss of interest	1006	1	5	2.71	1.354	0.266	0.077	-1.102	0.154
Friends versus money	1004	1	5	3.78	1.125	-0.606	0.077	-0.392	0.154
Funniness	1006	1	5	3.29	1.129	-0.191	0.077	-0.611	0.154
Fake	1009	1	5	2.13	1.047	0.823	0.077	0.118	0.154

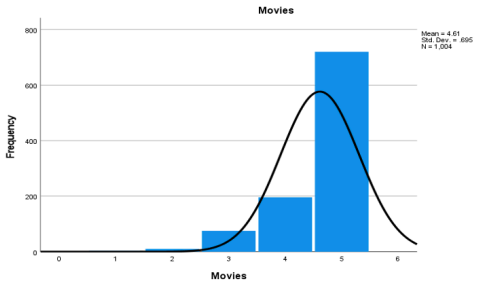
Criminal damage	1003	1	5	2.6	1.504	0.4	0.077	-1.311	0.154
Decision making	1006	1	5	3.2	1.201	-0.026	0.077	-0.858	0.154
Elections	1007	1	5	3.42	1.575	-0.43	0.077	-1.359	0.154
Self-criticism	1005	1	5	3.58	1.193	-0.474	0.077	-0.643	0.154
Judgment calls	1006	1	5	3.99	0.972	-0.665	0.077	-0.228	0.154
Hypochondria	1006	1	5	1.91	1.157	1.101	0.077	0.208	0.154
Empathy	1005	1	5	3.86	1.132	-0.774	0.077	-0.209	0.154
Eating to survive	1010	1	5	2.23	1.214	0.733	0.077	-0.454	0.154
Giving	1004	1	5	2.98	1.309	0.042	0.077	-1.041	0.154
Compassion to animals	1003	1	5	3.97	1.19	-0.885	0.077	-0.307	0.154
Borrowed stuff	1008	1	5	4.02	1.053	-0.967	0.077	0.288	0.154
Loneliness	1009	1	5	2.89	1.132	0.19	0.077	-0.609	0.154
Cheating in school	1006	1	5	3.74	1.254	-0.469	0.077	-1.081	0.154
Health	1009	1	5	3.25	1.075	-0.219	0.077	-0.364	0.154
Changing the past	1008	1	5	2.95	1.278	0.061	0.077	-0.991	0.154
God	1008	1	5	3.3	1.483	-0.289	0.077	-1.297	0.154
Dreams	1010	1	5	3.3	0.683	-0.025	0.077	1.003	0.154
Charity	1007	1	5	2.1	1.031	0.635	0.077	-0.241	0.154
Number of friends	1010	1	5	3.34	1.055	-0.122	0.077	-0.38	0.154
Waiting	1007	1	5	2.67	1.003	0.152	0.077	-0.353	0.154
New environment	1008	1	5	3.48	1.152	-0.402	0.077	-0.562	0.154
Mood swings	1006	1	5	3.26	1.045	0.072	0.077	-0.67	0.154

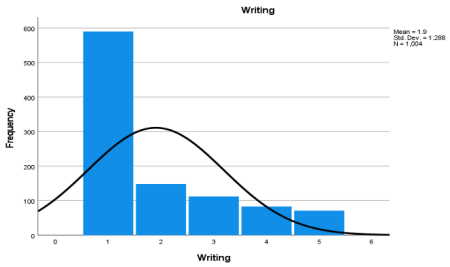
Appearence and gestures	1007	1	5	3.6	0.944	-0.351	0.077	-0.123	0.154
Socializing	1005	1	5	3.16	1.093	-0.069	0.077	-0.536	0.154
Achievements	1008	1	5	2.96	0.936	-0.095	0.077	-0.084	0.154
Responding to a serious letter	1004	1	5	3.07	1.172	-0.186	0.077	-0.74	0.154
Children	1006	1	5	3.62	1.121	-0.366	0.077	-0.676	0.154
Assertiveness	1008	1	5	3.52	1.103	-0.339	0.077	-0.558	0.154
Getting angry	1006	1	5	3.01	1.174	-0.047	0.077	-0.765	0.154
Knowing the right people	1008	1	5	3.49	1.092	-0.448	0.077	-0.313	0.154
Public speaking	1008	1	5	3.52	1.268	-0.435	0.077	-0.838	0.154
Unpopularity	1007	1	5	3.46	1.118	-0.353	0.077	-0.355	0.154
Life struggles	1007	1	5	3.03	1.375	-0.039	0.077	-1.226	0.154
Happiness in life	1006	1	5	3.71	0.824	-0.551	0.077	0.493	0.154
Energy levels	1005	1	5	3.63	1.002	-0.475	0.077	-0.133	0.154
Small - big dogs	1006	1	5	2.97	1.223	-0.053	0.077	-0.763	0.154
Personality	1006	1	5	3.29	0.643	0.113	0.077	1.302	0.154
Finding lost valuables	1006	1	5	2.87	1.244	0.048	0.077	-0.865	0.154
Getting up	1005	1	5	3.59	1.31	-0.5	0.077	-0.931	0.154
Interests or hobbies	1007	1	5	3.55	1.171	-0.36	0.077	-0.706	0.154
Parents' advice	1008	1	5	3.27	0.866	-0.348	0.077	0.253	0.154
Questionnaires or polls	1006	1	5	2.75	1.102	0.165	0.077	-0.467	0.154
Finances	1007	1	5	3.02	1.144	-0.154	0.077	-0.67	0.154
Shopping centres	1008	1	5	3.23	1.323	-0.212	0.077	-1.083	0.154

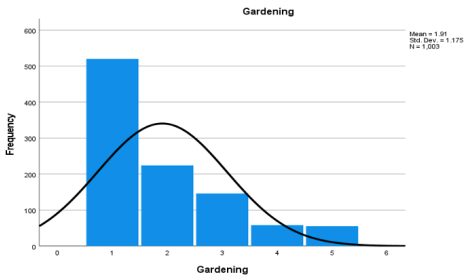
Branded clothing	1008	1	5	3.05	1.306	-0.142	0.077	-1.039	0.154
Entertainment spending	1007	1	5	3.2	1.189	-0.124	0.077	-0.829	0.154
Spending on looks	1007	1	5	3.11	1.205	-0.092	0.077	-0.876	0.154
Spending on gadgets	1010	1	5	2.87	1.285	0.181	0.077	-1.021	0.154
Spending on healthy eating	1008	1	5	3.56	1.094	-0.407	0.077	-0.54	0.154
Age	1003	15	30	20.43	2.829	1.149	0.077	1.579	0.154
Height	990	62	203	173.51	10.025	-1.152	0.078	14.556	0.155
Weight	990	41	165	66.41	13.84	1.224	0.078	3.969	0.155
Number of siblings	1004	0	10	1.3	1.013	1.769	0.077	7.435	0.154

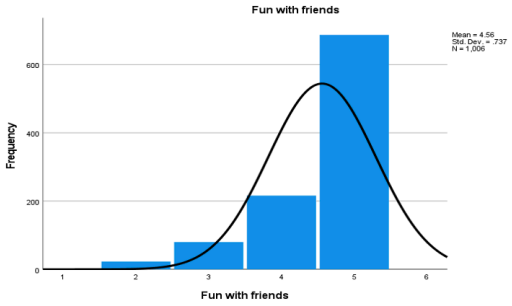
It is further validated by looking through the data frequency tables and plotting them visually on the following histogram, and those for the other variables that were found to be normally distributed are included under *Appendix B.*

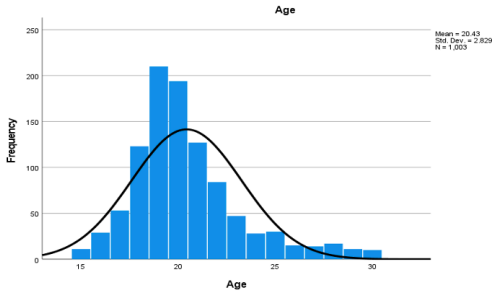


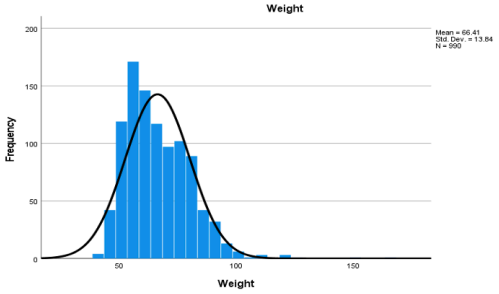


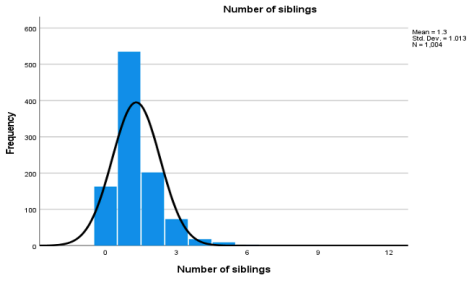












Correlation matrix were generated next to check for variables that might be highly correlated with each other, shown under *Appendix C*. Given the close relation between 'Biology' and 'Medicine', the two were found to be strongly associated with each other, with a Pearson correlation of 0.716. The rest of the variables were found to have poor and weak correlation.

3.2 Target Variable Derivation

Another critical step prior to modelling is finding which question would be utilized as the target variable. The data includes various questions relating to finance and spending habits, all of which could potentially be used as the dependent variable. *Shoppingcentres, Brandedclothing, Entertainmentspending, Spendingonlooks, Spendingongadgets* were aggregated to compute for an average spend rating, which is beneficial for ecommerce industry in targeting their customers.

A linear regression using the mean spend rating *(MeanSpendRating)* as dependent variable and all other variables as predictor was generated to gain initial insights on the model results. The model returned an adjusted r-square value of 0.372 with the following significant variables and coefficients.

Coefficientsa								
Model		**Unstandardized Coefficients**		**Standardized Coefficients**	**t**	**Sig.**	**Collinearity Statistics**	
		B	Std. Error	Beta			Tolerance	VIF
26	(Constant)	1.895	0.25		7.595	<.001		
	Knowing the right people	0.135	0.02	0.188	6.875	<.001	0.833	1.2
	Appearence and gestures	0.126	0.022	0.152	5.66	<.001	0.864	1.158
	Adrenaline sports	0.057	0.015	0.103	3.765	<.001	0.824	1.213
	Celebrities	0.071	0.017	0.115	4.122	<.001	0.798	1.253
	Finances	-0.097	0.018	-0.142	-5.393	<.001	0.897	1.115
	Internet	0.066	0.023	0.078	2.921	0.004	0.878	1.139
	Folk	-0.085	0.018	-0.124	-4.622	<.001	0.867	1.154
	Science and technology	0.068	0.018	0.111	3.818	<.001	0.738	1.356
	Health	0.069	0.02	0.094	3.44	<.001	0.832	1.202
	Loss of interest	0.045	0.015	0.077	2.939	0.003	0.905	1.105
	Economy Management	0.037	0.015	0.064	2.4	0.017	0.884	1.131
	Mood swings	-0.047	0.021	-0.062	-2.256	0.024	0.825	1.213
	Age	-0.024	0.007	-0.088	-3.356	<.001	0.904	1.106
	Documentary	-0.061	0.019	-0.088	-3.166	0.002	0.806	1.241
	Healthy eating	0.057	0.022	0.069	2.573	0.01	0.876	1.141
	alcohol_recode=drink a lot	0.152	0.05	0.08	3.019	0.003	0.879	1.137
	Giving	0.038	0.017	0.062	2.262	0.024	0.816	1.226
	gender_recode=female	-0.161	0.049	-0.101	-3.274	0.001	0.654	1.528
	Metal or Hardrock	-0.035	0.015	-0.06	-2.266	0.024	0.878	1.139
	Pets	0.029	0.013	0.058	2.218	0.027	0.911	1.098
	Charity	0.053	0.02	0.07	2.606	0.009	0.863	1.158
	Cheating in school	0.041	0.017	0.065	2.444	0.015	0.882	1.133

	Loneliness	-0.05	0.019	-0.072	-2.661	0.008	0.856	1.168
	Dancing	0.031	0.015	0.058	2.036	0.042	0.768	1.303
a Dependent Variable: MeanSpendRating								

Coefficients were low due to the nature of each scale variables. As an alternative, a binary variable to classify spenders (with 4 and above mean spend rating) and non-spenders (1 to 3 mean spend rating) were generated so that a logistic regression model can be ran.

3.3 Data Cleansing

No inaccuracies in the respondent-level data were found during the data exploration phase, hence no further modifications and removal were necessary pre-modelling.

3.4 Data Exploration Summary

A few scale questions were found to be skewed during data exploration; however, this result is quite common in scale measurement questions particularly for short ranges (1 to 5). Age, height, and weight were also skewed due to employing snowball sampling, resulting to close similarity in demographic profile among respondents. Apart from this, no data cleaning was required pre-modelling. Binary variable to classify spenders versus non-spenders were built based on specific spend questions from the data to allow running of logistic regression.

4. DATA PREPARATION AND FEATURE ENGINEERING

4.1 Data Preparation Needs

Variables with missing data need to be handled through imputation prior to modelling to refrain from samples being excluded from the regression model. Also, the number of variables need to be reduced by removing redundant questions or via dimension reduction techniques. Subsequent data processing and feature engineering were applied using Python.

4.2 Variable Transformations, Imputation, and Exclusions

Redundant and impractical variables need to be identified and excluded from the dataset. Spend questions that were used in determining the target variable, *spender vs non-spender*, need to be omitted from the list of predictor variables as well as shopping interests and finance habits, which may seem to be redundant. Questions related to phobia, left/ right hand dominance, lying, and punctuality, were also dropped from the analysis, as they are difficult and impractical in actual circumstances. The *onlychild* question is also removed in lieu of relating to number of siblings.

During the initial data exploration stage, it was found that height and weight were among highly skewed variables. Upon further investigation, there were indeed some outliers in these variables, and it might be beneficial to impose cap and floor their ranges.

```
##cap and floor Height and Weight variables
millenials2['Height'] = np.where(millenials2['Height'] <
148,148,millenials2['Height'])
millenials2['Weight'] = np.where(millenials2['Weight'] <
43,43,millenials2['Weight'])
millenials2['Weight'] = np.where(millenials2['Weight'] >
100,100,millenials2['Weight'])
```

Response counts for doctorate degree was found to be small during the data exploration. To make all levels of education robust, variable was recoded to combine some segments together.

```
#recoding eduction to reduce segments with small counts
millenials2['Education'] =
millenials2['Education'].map({"college/bachelor degree":
"college/bachelor degree",
       "currently a primary school pupil": "Primary school",
       "primary school": "Primary school",
       "doctorate degree": "post-grad",
       "masters degree": "post-grad",
       "secondary school": "secondary school"})
```

'*InternetUsage'* and '*Smoking'* variable was also made as binary for robustness.

```
#recode Internet usage
millenials2['Internetusage'] = millenials2['Internetusage'].map(
    {
    "few hours a day": "few hours a day",
    "less than an hour a day":"less than an hour a day",
    "most of the day": "most of the day",
    "no time at all":"less than an hour a day"
    })

#recode Smoker/Non-Smoker
millenials2['Smoking'] = millenials2['Smoking'].map(
    {
    "current smoker": "smoker",
    "former smoker" : "smoker",
    "never smoked"  : "non-smoker",
    "tried smoking" : "non-smoker"
    })
```

While minute, some of the variables having 1 to 2% missing cases needs to be imputed to be feasible for regression. Mean and mode were used for replacing numeric variables and categorical, respectively, as shown in the following Python codes.

```
#imputing numeric variables with mean
c = millenials2.select_dtypes(np.number).columns
millenials2[c] = millenials2[c].fillna(millenials2[c].mean())

#imputing categorical variables with mode (use original categorical
variables for ease of generating dummy for regression)
millenials2['Smoking']=millenials2['Smoking'].fillna(millenials2['Smoki
ng'].mode()[0])
millenials2['Alcohol']=millenials2['Alcohol'].fillna(millenials2['Alcoh
ol'].mode()[0])
millenials2['Punctuality']=millenials2['Punctuality'].fillna(millenials
2['Punctuality'].mode()[0])
millenials2['Lying']=millenials2['Lying'].fillna(millenials2['Lying'].m
ode()[0])
millenials2['Internetusage']=millenials2['Internetusage'].fillna(millen
ials2['Internetusage'].mode()[0])
millenials2['Gender']=millenials2['Gender'].fillna(millenials2['Gender'
].mode()[0])
millenials2['Education']=millenials2['Education'].fillna(millenials2['E
ducation'].mode()[0])
```

```
millenials2['Villagetown']=millenials2['Villagetown'].fillna(millenials
2['Villagetown'].mode()[0])
millenials2['Houseblockofflats']=millenials2['Houseblockofflats'].filln
a(millenials2['Houseblockofflats'].mode()[0])
```

Dimension reduction variable inputs were also standardized prior to make scale consistent across.

```
from sklearn.preprocessing import StandardScaler

#standardize the data

scaler = StandardScaler()

scaled_psych = psych_df
scaled_psych = pd.DataFrame(scaler.fit_transform(psych_df),
columns=scaled_psych.columns)
```

4.3 Feature Engineering

The dataset 117 variables covering a person's preferences, hobbies, interests, and personality traits; hence, it can be supposed that there are hidden or latent variables that connect some of these variables together. A multivariate technique called Exploratory Factor Analysis to determine these latent variables and their grouping. Instead of using the individual variables as inputs in the model, the factors would be used instead for model simplification.

Bartlett's test and KMO were generated to verify that the dataset is suitable for Factor Analysis. Bartlett's test resulted to a low p-value, concluding that there are differences in variances across variables and a strong KMO statistics supported the assumption that these variances are likely caused by underlying factors.

```
#CHECK ADEQUACY

#Bartlett
#p-value should be less than 0.05 (statistically sig.)
chi_square_value,p_value=calculate_bartlett_sphericity(matrix)
print('Chi-square: ', chi_square_value, 'P-Value: ', p_value)
```

```
#KMO
#Value should be > than 0.6
kmo_all,kmo_model=calculate_kmo(matrix)
print('KMO: ', kmo_model)
```

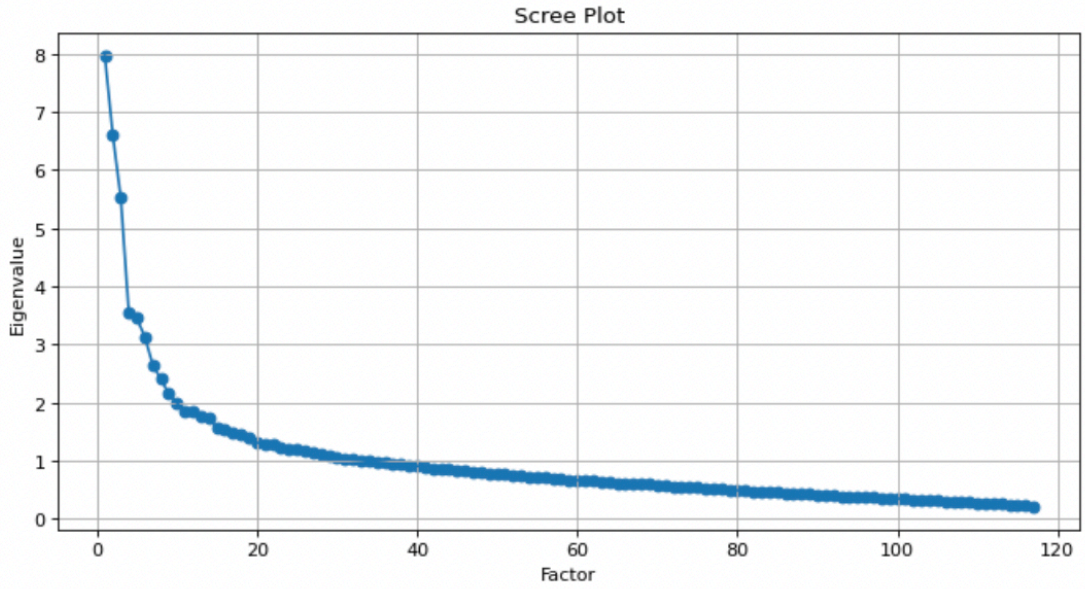
```
Chi-square:  14392.60605665464 P-Value:  0.0
KMO:  0.893632795537232
```

The number of factors were determined by evaluating the eigenvalues and scree plot.

```
#get the eigenvectors and eigenvalues

ev, v = fa.get_eigenvalues()
print(ev)
```

```
#do a scree plot

xvals = range(1,scaled_psych.shape[1]+1)
plt.figure(figsize=(10, 5), dpi=80)
plt.scatter(xvals,ev)
plt.plot(xvals, ev)
plt.title('Scree Plot')
plt.xlabel('Factor')
plt.ylabel('Eigenvalue')
plt.grid()
plt.show()
```

```
[7.97516438 6.60872968 5.5226094  3.54919607 3.45382305 3.13173904
 2.64932755 2.40737073 2.14922855 1.9930851  1.85662521 1.84079575
 1.76085033 1.72502694 1.56824434 1.52818088 1.47648996 1.45085709
 1.40901096 1.3094648  1.2975989  1.27097243 1.24182967 1.19806982
 1.19446411 1.16261094 1.13523228 1.12521721 1.08858368 1.06567426
 1.03442764 1.01976993 1.00981795 0.98831261 0.97931753 0.96529441
 0.9380393  0.93454651 0.92688171 0.91398553 0.87941581 0.87135796
 0.85847984 0.84940798 0.84301446 0.82691523 0.80051235 0.79297731
 0.78615551 0.77731355 0.7665851  0.75481494 0.743746   0.72512945
 0.72274419 0.71308663 0.7044423  0.68713784 0.67378622 0.665703
 0.6617787  0.65584568 0.6459647  0.63164642 0.61682342 0.61432635
 0.60526291 0.59718927 0.59366452 0.57923426 0.57258036 0.56232931
 0.558758   0.54918687 0.53563463 0.53180336 0.52672909 0.51544467
 0.5060829  0.4920982  0.48995232 0.47780081 0.47288902 0.46344387
 0.45761642 0.44480558 0.44002531 0.43550794 0.42994066 0.41801388
 0.41242557 0.40214461 0.39169687 0.38684799 0.38420386 0.3755432
 0.37295617 0.36200396 0.35745596 0.35212533 0.34176969 0.33679187
 0.32849578 0.32457025 0.31426898 0.30407685 0.30187809 0.29468832
 0.28581015 0.27490565 0.27432439 0.26648203 0.25981707 0.24163203
 0.23563619 0.22734421 0.20853363]
```



Recommended number of factors appears to be 36 based on above output (eigenvalues greater than 1 and arm bend for the scree plot). Variables with low communalities (less than 0.4) as well as those found to have poor loading across all factors were also dropped prior to running factor models.

```
#removing variables with low communalities
scaled_psych2= scaled_psych.drop(columns=['Music',
'Slowsongsorfastsongs',
'Comedy',
'Romantic',
'Psychology',
'Reading',
'Cars',
```

```
'Countrysideoutdoors',
'Musicalinstruments',
'Gardening',
'Celebrities',
'Healthyeating',
'Lossofinterest',
'Criminaldamage',
'Eatingtosurvive',
'Giving',
'Borrowedstuff',
'Cheatinginschool',
'Health',
'Dreams',
'Waiting',
'Achievements',
'Respondingtoaseriousletter',
'Children',
'Assertiveness',
'Publicspeaking',
'Lifestruggles',
'Personality',
'Gettingup',
'Questionnairesorpolls',])
#scaled_psych2
```

Various factor analysis solutions (principal axis, maximum likelihood, and minimum residual) and rotation techniques (varimax, equamax, and promax) as parameters were explored to determine the optimal factor groupings, using FactorAnalyzer code package below.

```
fa = FactorAnalyzer(n_factors = 36, method='minres', rotation='promax')
fa.fit(scaled_psych)
```

Generating 36 factors resulted to many single-item factors, hence it would be advisable to check results based on a smaller number of factors. Using 32 factors provided a more robust result. A summary comparison of different solutions and rotation are provided in the following table.

FA Summary (After dropping low communalities and no loading)					
	PF - Varimax	Minres - Varimax	ML - Varimax	Minres - Promax	ML - Promax
Total Variance Explained	67%	53%	55%	74%	67%
% Unique Loading	72%	85%	76%	82%	84%

Using minimum residual solution with Promax rotation gave the best total variance explained, and 82% of the variables are loaded uniquely to a single factor; hence, this output will be used for dimension reduction. Promax rotation is also commonly used in most psychology studies where relationship among variables is assumed.

In summary, there are 32 underlying factors among the 117 input variables on preferences, interests, and personality traits. Factor loadings and naming are shown in *Appendix D.*

4.4 Factor Scores Computation

Though more refined techniques are available in computing factor scores, a simpler sum of scores or averages are preferred due to their ease of interpretation and practicality. Factors corresponding to negative loadings, such as *loneliness* and *Changingthepast* which has negative effect on an individual's optimistic attitude, were reversed so that all scores would yield as positive. After which, each factor was computed using mean of the variables that are contributing the high loadings.

```
old = [1,2,3,4,5]
new = [5,4,3,2,1]

millenials2['Loneliness'] = millenials['Loneliness'].replace(old, new)
millenials2['Changingthepast'] =
millenials['Changingthepast'].replace(old, new)
```

```
#Factor scores computation using mean

millenials2['Art']=(millenials2['Musical']+millenials2['Artexhibitions'
]+millenials2['Theatre']+millenials2['Opera']+millenials2['Writing']+mi
llenials2['Classicalmusic'])/6
millenials2['Sciences']=(millenials2['Medicine']+millenials2['Biology']
+millenials2['Chemistry'])/3
millenials2['RockMusic']=(millenials2['Rock']+millenials2['MetalorHardr
ock']+millenials2['Punk'])/3
millenials2['Happiness']=(millenials2['Energylevels']+millenials2['Lone
liness']+millenials2['Changingthepast']+millenials2['Happinessinlife'])
/4
millenials2['Educational']=(millenials2['Documentary']+millenials2['His
tory'])/2
millenials2['Spritualbeliefs']=(millenials2['God']+millenials2['Finalju
dgement']+millenials2['Religion'])/3
millenials2['Physicalactivity']=(millenials2['Activesport']+millenials2
['Interestsorhobbies']+millenials2['Adrenalinesports'])/3
millenials2['PopLatino']=(millenials2['Latino']+millenials2['Pop'])/2
millenials2['Countrymusic']=(millenials2['Country']+millenials2['Folk']
)/2
millenials2['Socialawareness']=(millenials2['Law']+millenials2['Economy
Management'])/2
millenials2['Computertechnology']=(millenials2['Internet']+millenials2[
'PC'])/2
millenials2['Fantasy']=(millenials2['FantasyFairytales']+millenials2['A
nimated'])/2
millenials2['Extrovert']=(millenials2['Socializing']+millenials2['Newen
vironment'])/2
millenials2['ThrillerHorror']=(millenials2['Thriller']+millenials2['Hor
ror'])/2
millenials2['Upbeatmusic']=(millenials2['Dance']+millenials2['TechnoTra
nce']+millenials2['Dancing'])/3
millenials2['Rnb']=(millenials2['SwingJazz']+millenials2['Rocknroll'])/
2
millenials2['Diligence']=(millenials2['Writingnotes']+millenials2['Prio
ritisingworkload']+millenials2['Workaholism'])/3
millenials2['MathAndPhysics']=(millenials2['Physics']+millenials2['Math
ematics'])/2
millenials2['HipMusic']=(millenials2['ReggaeSka']+millenials2['HiphopRa
p'])/2
millenials2['Trendy']=(millenials2['Appearenceandgestures']+millenials2
['Dailyevents']+millenials2['Knowingtherightpeople'])/3
millenials2['Friendliness']=(millenials2['Funwithfriends']+millenials['
Friendsversusmoney']+millenials2['Numberoffriends'])/3
millenials2['Trustworthiness']=(millenials2['Keepingpromises']+millenia
ls2['Reliability'])/2
```

```
millenials2['Loveforanimals']=(millenials2['Compassiontoanimals']+mille
nials2['Pets'])/3
millenials2['Irritability']=(millenials2['Moodswings']+millenials2['Get
tingangry'])/2
millenials2['Pretentiousness']=(millenials2['Fake']+millenials2['Funnin
ess'])/2
millenials2['Conscientiousness']=(millenials2['Selfcriticism']+millenia
ls2['Thinkingahead']+millenials2['Decisionmaking'])/3
millenials2['WesternWarFilms']=(millenials2['War']+millenials2['Western
'])/2
millenials2['ScifiActionMovie']=(millenials2['Action']+millenials2['Sci
fi'])/2
millenials2['Righteousness']=(millenials2['Unpopularity']+millenials2['
Parentsadvice']+millenials2['Findinglostvaluables'])/3
millenials2['PoliticalAwareness']=(millenials2['Politics']+millenials2[
'Elections'])/2
millenials2['LanguageGeography']=(millenials2['Geography']+millenials2[
'Foreignlanguages'])/2
millenials2['Empathy2']=(millenials2['Judgmentcalls']+millenials2['Empa
thy']+millenials2['Charity'])/3
```

18 out of the 32 factors on personality were deemed to be likely to affect spending and worth including in the model. These variables would be included together with other demographic and categorical variables as input or predictor variables in classifying Gen Z and millennial spenders.

5. MODEL EXPLORATION

5.1 Modeling Approach

Classification techniques namely Decision Tree and Logistic Regression were employed, with binary classification *high spender versus low spender* as the target variable and the psychographic factors from the Exploratory Factor Analysis results together with demographic variables (gender, education, height, weight) and other categorical variables (smoking, alcohol consumption, and internet usage) as predictor variables. Dataset was split equally between training and test datasets, where the former is used in model building while the latter is for results validation.

```
from sklearn.model_selection import train_test_split


X = millenials2[['Art', 'Happiness' , 'Spritualbeliefs',
'Physicalactivity', 'Socialawareness', 'Computertechnology',
'Extrovert', 'Diligence', 'Trendy', 'Friendliness', 'Trustworthiness',
'Loveforanimals', 'Irritability', 'Pretentiousness',
'Conscientiousness', 'Righteousness', 'PoliticalAwareness',
'Empathy2', 'Age', 'Height', 'Weight',  'Numberofsiblings', 'Smoking',
'Alcohol', 'Internetusage', 'Gender',  'Education', 'Villagetown',
'Houseblockofflats']]
y = millenials2['Spend_Binary']


train_X, valid_X, train_y, valid_y = train_test_split(X, y, test_size =
0.5, random_state=1)
```

The best model among the decision tree and logistic regression yielding the best accuracy based on the least average squared error (ASE) and least false positivity rate on the validation dataset will be determined and used, and the significant variables will be assessed and further applied in segmenting the respondents to come up with customer targeting.

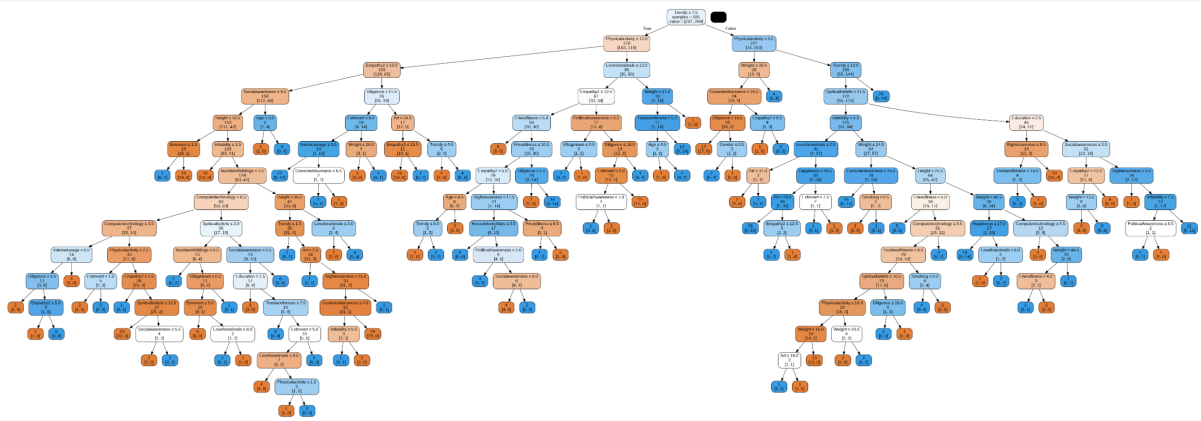
5.2 Decision Tree

'DecisionTreeClassifier' from sklearn library in Python were utilized in running the decision tree model. Required preprocessing to convert the string variables to numeric was also ran to make the dataset suitable for the Decision Tree package.

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import cross_val_score, GridSearchCV
import matplotlib.pylab as plt
!pip install dmba
from dmba import plotDecisionTree, classificationSummary


#converting string variables to numeric for decision tree
le = preprocessing.LabelEncoder()
train_X_DT = train_X.apply(le.fit_transform)
valid_X_DT = valid_X.apply(le.fit_transform)
```

A full-depth decision tree was initially generated and analyzed to check performance if all variables were to be fitted in the model.

```
fullClassTree = DecisionTreeClassifier(random_state=1)
fullClassTree.fit(train_X_DT,train_y)


plotDecisionTree(fullClassTree, feature_names=train_X_DT.columns)


#checking for accuracy using valid dataset
classificationSummary(valid_y,fullClassTree.predict(valid_X_DT))
```

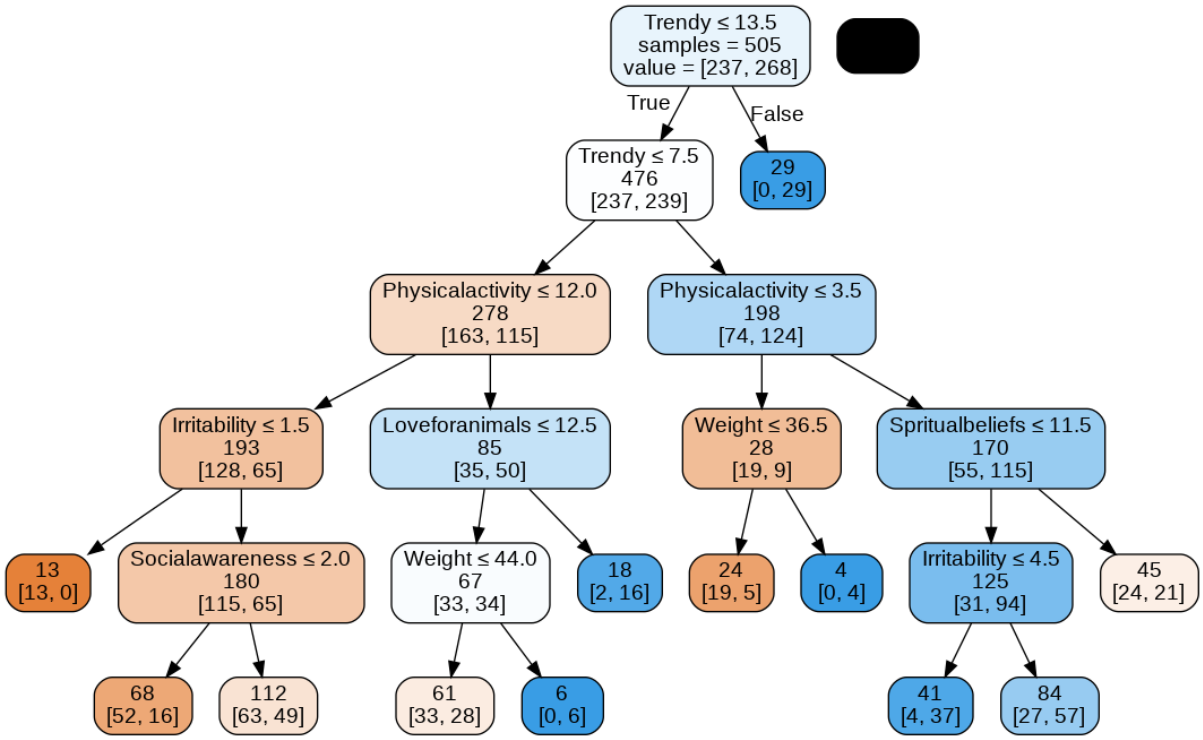


```
Confusion Matrix (Accuracy 0.5822)

       Prediction
Actual   0   1
     0 121 118
     1  93 173
```

Based on the full decision tree output, *Trendy, Physical activity, and love for animals* were the top variables having a strong impact on spending habits. Being trendy factor spans taking efforts to look good physically and being up to date to latest events and fads, which might raise individual's shopping inclinations. Physical activity likely increases a person's spending on more activities, sports apparel, and accessories while being compassionate with animals could drive one to opt for more sustainable and responsible products, which are usually more expensive. While overall accuracy seems to be acceptable at 58%, the false positivity rate seems to be

a bit high at 49%. Pruning could potentially improve the model performance. The most optimal tree depth and splits could be explored using grid search feature.

```
param_grid = {'criterion':['gini','entropy'],
    'max_depth':[5,7,8,9,10,11,12],
    'min_samples_split':[0.10,0.07,0.05,0.01,0.005],
    'min_impurity_decrease': [0.10,0.05, 0.02, 0.01, 0.001]
}


gridsearch = GridSearchCV(DecisionTreeClassifier(random_state=1),
param_grid, cv=5, n_jobs=1)
gridsearch.fit(train_X_DT, train_y)


gridsearch.best_score_


gridsearch.best_params_

gridClassTree =
DecisionTreeClassifier(criterion='entropy',max_depth=7,min_impurity_dec
rease=0.01,min_samples_split=0.05, random_state=1)
gridClassTree.fit(train_X_DT, train_y)
plotDecisionTree(gridClassTree, feature_names=train_X_DT.columns)

besttree=gridsearch.best_estimator_

plotDecisionTree(besttree, feature_names=train_X_DT.columns)

#checking for accuracy using valid dataset
classificationSummary(valid_y,besttree.predict(valid_X_DT))
```




```
Confusion Matrix (Accuracy 0.6257)

       Prediction
Actual   0   1
     0 178  61
     1 128 138
```

Optimal tree yielded depth of 7. Aside from being trendy, being physical active, and having love for animals, other variables that were found to be significant were weight, spiritual beliefs, and social awareness. Individuals with fitter weight are more likely to spend more on healthier options while those with stronger spiritual beliefs might tend to have less attachments and lower spending. When it comes to personality traits, people who are more irritable and less socially aware could turn to increased spending, possibly to cope. The model accuracy improved after pruning from 58% to 63%, and false positivity dropped to 26%.

5.3 Logistic Regression

'LogisticRegression' from sklearn library in python were utilized in running the logistic regression model. Categorical variables were converted into dummy variables for better result interpretation. A total of 33 variables, including dummy variables were then used as input variables in this model.

```
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score as acc

!pip install dmba
from dmba import classificationSummary, gainsChart, liftChart

millenials_df = millenials2[['Art', 'Happiness', 'Spritualbeliefs',
'Physicalactivity', 'Socialawareness', 'Computertechnology',
'Extrovert', 'Diligence', 'Trendy', 'Friendliness', 'Trustworthiness',
'Loveforanimals', 'Irritability','Pretentiousness','Conscientiousness',
'Righteousness', 'PoliticalAwareness', 'Empathy2', 'Age', 'Height',
'Weight', 'Numberofsiblings', 'Smoking', 'Alcohol', 'Internetusage',
'Gender', 'Education', 'Villagetown', 'Houseblockofflats',
'Spend_Binary']]


excluded_columns = 'Spend_Binary'
outcome = 'Spend_Binary'
predictors = [s for s in millenials_df.columns if s not in
excluded_columns]

#converting categorical variables to dummy variables
X = pd.get_dummies(millenials_df[predictors], drop_first=True)
y = millenials_df[outcome]
```

A full logistic regression classifier was initially generated, model was fitted using the training data where coefficients and odds ratio were obtained for output interpretation.

The python code used in the full logistic regression is found below.

```
millenial_logitreg = LogisticRegression(solver='liblinear', C=1e42,
random_state=1)
millenial_logitreg.fit(train_X, train_y)
```

```
full = pd.DataFrame({'coef': millenial_logitreg.coef_[0], 'odds':
np.e**millenial_logitreg.coef_[0]}, index=X.columns)
full
```

	coef	odds
Art	0.086351	1.090189
Happiness	0.333644	1.396046
Spritualbeliefs	-0.018486	0.981684
Physicalactivity	0.310614	1.364263
Socialawareness	0.277671	1.320052
Computertechnology	0.015934	1.016061
Extrovert	-0.019051	0.981129
Diligence	-0.157701	0.854105
Trendy	0.821874	2.274758
Friendliness	-0.063479	0.938494
Trustworthiness	0.255783	1.291473
Loveforanimals	0.461333	1.586187
Irritability	0.083389	1.086965
Pretentiousness	0.206948	1.229918
Conscientiousness	-0.003293	0.996713
Righteousness	0.092241	1.096629
PoliticalAwareness	-0.136755	0.872184
Empathy2	-0.023478	0.976796
Age	-0.076640	0.926223

Height	0.024984	1.025299
Weight	-0.000576	0.999424
Numberofsiblings	-0.100456	0.904425
Smoking_smoker	0.133576	1.142908
Alcohol_never	-0.615779	0.540220
Alcohol_social drinker	-0.142936	0.866810
Internetusage_less than an hour a day	-0.561504	0.570350
Internetusage_most of the day	0.558752	1.748490
Gender_male	-0.013839	0.986256
Education_college/bachelor degree	-0.282709	0.753739
Education_post-grad	-0.063375	0.938591
Education_secondary school	-0.072818	0.929770
Villagetown_village	-0.330454	0.718598
Houseblockofflats_house/bungalow	0.413523	1.512136

```
#confusion matrix using training data
classificationSummary(train_y, millenial_logitreg.predict(train_X))
```

```
Confusion Matrix (Accuracy 0.6792)

       Prediction
Actual   0   1
     0 147  90
     1  72 196
```

```
#confusion matrix using validation data
classificationSummary(valid_y, logitreg_pred)
```

```
Confusion Matrix (Accuracy 0.6218)

       Prediction
Actual   0   1
     0 146  93
     1  98 168
```

Based on the Full Logistic Regression output - *Trendy*, *Internetusage_most of the day*, *Loveforanimals*, *Houseblockoffflats_house/bungalow*, *Happiness, Physicalactivity, Socialawareness* are some of the top features found to have strong impact on spending habits. Based on validation data, overall classification accuracy is at 62% and the false positivity rate at 39%.

Given the vast set of features used in the full regression, some variables may be irrelevant or less significant to the target variable. Because of that, the process of *Feature Selection* was employed to reduce the number of input variables to include only those that will be most useful in predicting the target variable. *Forward Feature Selection*, *Backward Feature Selection* and *Stepwise Feature Selection* were each run separately. The different models were fitted using training data where the relevant features selected were utilized as predictors. The next step was to obtain the respective coefficients and odds ratio for model interpretability.

The *'SequentialFeatureSelection'* in mlxtend in python was applied as it has the parameters that allows defining the different feature selection methods. As a first step,

Logistic Regression classifier was built for usage in the feature selection. Next, feature selection was then created by defining the parameters relevant to the *forward*, *backward* and *stepwise* selection models. The parameters included are: *k_features*, *forward*, *floating*, *verbose*, *scoring*, and *cv*. The feature selection models were then fitted using training data to obtain the most significant variables. Finally, logistic regression model was applied on the selected features based on each of the selection models.

Forward Feature Selection

The forward feature selection starts with a null model and starts with the evaluation of each independent variable in the dataset one at a time. It selects variables that best performs in the model. All possible combinations of the selected variable and the subsequent features are then evaluated until the 'best' features are selected.

The python code used in the forward feature selection model is found below.

```
!pip install mlxtend
import joblib
import sys
sys.modules['sklearn.externals.joblib'] = joblib
from mlxtend.feature_selection import SequentialFeatureSelector as sfs

# Build RF classifier to use in feature selection
clf = LogisticRegression(solver='liblinear', C=1e42, random_state=1)

# Build step forward feature selection
sfs1 = sfs(clf,
           k_features='best',
           forward=True,
           floating=False,
           verbose=2,
           scoring='accuracy',
           cv=5)
```

```
# Perform the forward selection using train data
sfs1 = sfs1.fit(train_X, train_y)

# Which features?
feat_cols = list(sfs1.k_feature_idx_)


# Build regression model with selected features
clf = LogisticRegression(solver='liblinear', C=1e42, random_state=1)
clf.fit(train_X.iloc[:, feat_cols], train_y)

#calculate predictions using predict
y_train_predfwd = clf.predict(train_X.iloc[:, feat_cols])
y_valid_predfwd = clf.predict(valid_X.iloc[:, feat_cols])

#calculate probabilities using predict_proba
logitreg_probfwd = clf.predict_proba(valid_X.iloc[:,feat_cols])

#display the selected features, coefficients and odds ratio
display('intercept', clf.intercept_)
forward = pd.DataFrame({'coef': clf.coef_[0], 'odds':
np.e**clf.coef_[0]}, index=train_X.iloc[:, feat_cols].columns)
forward
```

	coef	odds
Art	0.010365	1.010419
Physicalactivity	0.343614	1.410035
Computertechnology	0.075658	1.078594
Extrovert	0.021139	1.021364
Trendy	0.885849	2.425043
Friendliness	-0.032563	0.967962
Trustworthiness	0.152942	1.165257
Loveforanimals	0.397471	1.488057
Conscientiousness	-0.069710	0.932665
PoliticalAwareness	-0.006527	0.993495
Gender_male	0.249165	1.282953
Education_college/bachelor degree	-0.174190	0.840137

```
#confusion matrix using training data
classificationSummary(train_y, y_train_predfwd)
```

```
Confusion Matrix (Accuracy 0.6713)

       Prediction
Actual   0   1
     0 151  86
     1  80 188
```

```
#confusion matrix using validation data
classificationSummary(valid_y, y_valid_predfwd)
```

```
Confusion Matrix (Accuracy 0.6515)

       Prediction
Actual   0   1
     0 151  88
     1  88 178
```

Out of the initial 33 variables, 12 variables were deemed significant after running the forward feature selection model. A slight improvement on the classification accuracy was observed from 62% to 65%. Conversely, the false positivity rate dropped by 2% which is seen to be an improvement in the model performance.

Backward Elimination Feature Selection

The backward elimination on the other hand starts with the entire set of features and works in reverse. The algorithm then removes the variables found to be irrelevant and only keeping the 'best' variables in the model.

The python code used in the backward elimination feature selection is found below.

```
# step backward elimination

clf3 = LogisticRegression(solver='liblinear', C=1e42, random_state=1)

sfs2 = sfs(clf3,
           k_features='best',
```

```
            forward=False,
            floating=False,
            verbose=2,
            scoring='accuracy',
            cv=5)

sfs2 = sfs2.fit(np.array(train_X), train_y)


# Which features?
feat_cols2 = list(sfs2.k_feature_idx_)

# Build regression model with selected features
clf3 = LogisticRegression(solver='liblinear', C=1e42, random_state=1)
clf3.fit(train_X.iloc[:, feat_cols2], train_y)

#calculate predictions using predict
y_train_predbwd = clf3.predict(train_X.iloc[:, feat_cols2])
y_valid_predbwd = clf3.predict(valid_X.iloc[:, feat_cols2])

#calculate probabilities using predict_proba
logitreg_probbwd = clf3.predict_proba(valid_X.iloc[:,feat_cols2])

#display the selected features, coefficients and odds ratio
display('intercept', clf3.intercept_)
display(pd.DataFrame({'coef': clf3.coef_[0], 'odds':
np.e**clf3.coef_[0]}, index=train_X.iloc[:, feat_cols2].columns))
```

	coef	odds
Physicalactivity	0.378139	1.459566
Diligence	-0.155093	0.856335
Trendy	0.866104	2.377629
Trustworthiness	0.197453	1.218296
Loveforanimals	0.421742	1.524615
Alcohol_never	-0.458821	0.632028
Internetusage_less than an hour a day	-0.687613	0.502775

```
#confusion matrix using training data
classificationSummary(train_y, y_train_predbwd)
```

```
Confusion Matrix (Accuracy 0.6614)

       Prediction
Actual   0   1
     0 142  95
     1  76 192
```

```
#confusion matrix using validation data
classificationSummary(valid_y, y_valid_predbwd)
```

```
Confusion Matrix (Accuracy 0.6416)

       Prediction
Actual   0   1
     0 145  94
     1  87 179
```

The backward elimination model resulted to 7 significant features from the initial 33 features. The model's classification accuracy also slightly improved from 62% to 64%. However, no improvement was observed in its false positivity rate as it remained to be at 39%.

Stepwise Feature Selection

The stepwise feature selection is a combination of the forward feature selection and backward elimination. Such that while adding a new variable (forward), it at the same time checks the significance of the earlier selected variables and removes the variables found to be irrelevant (backward elimination).

The python code used in the stepwise feature selection is found below.

```
# stepwise feature elimination

clf5 = LogisticRegression(solver='liblinear', C=1e42, random_state=1)
sfs3 = sfs(clf5,
           k_features='best',
```

```
            forward=True,
            floating=True,
            verbose=2,
            scoring='accuracy',
            cv=5)

sfs3 = sfs3.fit(np.array(train_X), train_y)

# Which features?
feat_cols3 = list(sfs3.k_feature_idx_)

# Build regression model with selected features
clf5 = LogisticRegression(solver='liblinear', C=1e42, random_state=1)
clf5.fit(train_X.iloc[:, feat_cols3], train_y)

#calculate predictions using predict
y_train_predsw = clf5.predict(train_X.iloc[:, feat_cols3])
y_valid_predsw = clf5.predict(valid_X.iloc[:, feat_cols3])

#calculate probabilities using predict_proba
logitreg_probsw = clf5.predict_proba(valid_X.iloc[:,feat_cols3])

#display the selected features, coefficients and odds ratio
display('intercept', clf5.intercept_)
display(pd.DataFrame({'coef': clf5.coef_[0], 'odds':
np.e**clf5.coef_[0]}, index=train_X.iloc[:, feat_cols3].columns))
```

	coef	odds
Art	0.008347	1.008382
Physicalactivity	0.342993	1.409159
Computertechnology	0.076032	1.078997
Extrovert	0.021004	1.021226
Trendy	0.883496	2.419342
Friendliness	-0.032584	0.967941
Trustworthiness	0.152426	1.164656
Loveforanimals	0.397512	1.488118
Conscientiousness	-0.070413	0.932009
Gender_male	0.246457	1.279484
Education_college/bachelor degree	-0.175770	0.838811

```
#confusion matrix using training data
classificationSummary(train_y, y_train_predsw)
```

```
Confusion Matrix (Accuracy 0.6713)

       Prediction
Actual   0   1
     0 150  87
     1  79 189
```

```
#confusion matrix using validation data
classificationSummary(valid_y, y_valid_predsw)
```

```
Confusion Matrix (Accuracy 0.6495)

       Prediction
Actual   0   1
     0 151  88
     1  89 177
```

The stepwise feature selection found 11 significant variables deemed to have a strong influence on spending habits. The model's classification accuracy improved from 62% to 65%, whereas an improvement in the false positivity rate was also observe after it dropped by 2%

6.0 MODEL RECOMMENDATION

6.1 Model Selection

The classification report for the Decision Tree and Logistic Regression models is summarized in the following table. The computation of the relevant metrics is found in *Appendix E to Appendix J.*.

	Decision Tree		Logistic Regression			
Validation	**Full DT**	**Grid Search**	**Full**	**Forward**	**Backward**	**Stepwise**
Classifier Accuracy	58.22%	62.57%	62.18%	65.15%	64.14%	64.95%
Recall / Sensitivity (TPR)	65.04%	51.88%	63.16%	66.92%	67.29%	66.54%
Specificity / Selectivity (TNR)	50.63%	74.48%	61.09%	63.18%	60.67%	63.18%
False Positive Rate (FPR)	49.37%	25.52%	38.91%	36.82%	39.33%	36.82%
False Negative Rate (FNR)	34.96%	48.12%	36.84%	33.08%	32.71%	33.46%
Precision / Positive Predictive Value	59.45%	69.35%	64.37%	66.92%	65.57%	66.79%
F1 - Score	62.12%	59.35%	63.76%	66.92%	66.42%	66.67%
Average Squared Error (ASE)	0.42	0.37	0.38	0.35	0.36	0.35
AUC	0.58	0.67	0.68	0.69	0.69	0.69

Based on the several models applied, the best model among the decision tree and logistic regression, the *Forward Selection* and the *Stepwise Selection* models yielded the best accuracy based on the least average squared error (ASE) and least false positivity rate on the validation dataset. Both models achieved a classification accuracy of 65%, and false positivity rates of 37% and average squared errors (ASE) of 0.35.

6.2 Model Theory

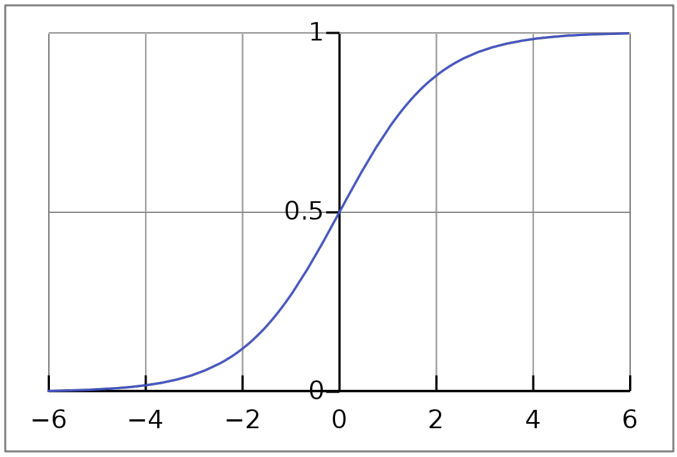
Decision Tree and Logistic Regression classification algorithms were employed for this project given the binary target variable outcome as *high spender (1) versus low spender (0)*.

i. **Decision Trees** are non-parametric supervised learning technique popularly used in classification and regression problems. Its main aim is to build a model that predicts the value of an outcome or a target variable through learning decision rules gathered from the independent data variables or features. Among all other algorithms, Decision Trees is one of the simplest classification method due to its straightforward approach and ease of model interpretability.

Decision tree splits the various set of variables into several and smaller subsets where data points continuously split until a related decision tree is further developed. Its final output consists of a decision node and leaf nodes wherein decision nodes may consist of two or more branches. The leaf node, on the other hand, corresponds to a classification or a decision where the topmost decision node found in in the tree refers to the best predictor in the model, or also known as the root node.

ii. **Logistic Regression** algorithm also belong to the group of supervised learning and is similarly used when the dependent or target variable is categorical. For this project, the binary logistic regression was employed as there are only 2 possible outcomes. The model estimates the probability of an outcome, a high spender (1) or a low spender (0) based on a vast set of independent variables or features.

In logistic regression, logit transformation is employed on the odds which refers to the probability of success (1) versus the probability of failure (0), or also known as the log odds. The logistic function is represented by the following formula: $Logit\ (x) = \frac{1}{1+e^{-x}}$



While Decision Trees are considered to be one of the simplest methods in terms of classification, they are more likely to overfit the data because they split on several different combination of features. Although Logistic Regression is less inclined to overfitting, such scenarios may still be observed especially in high dimensional datasets but may be avoided through regularization or standardization techniques.

For this project, logistic regression was preferred, not only because it yielded the best accuracy based on the least average squared error (ASE) and least false positivity rate but also allows better interpretation of the model coefficients explaining the features or the variables significance to the target variable.

Ultimately, the theory on buyer behavior was also inferred based on the output derived from applying the logistic regression algorithm since it is line with the objective of this project. The buyer behavior theory, also known as the "Howard-Sheth Model of Consumer Behavior" was established in 1969 by John Howard and Jagadish Sheth. The objective of this theory focuses on explaining the impact of different factors such as social, psychological, influences regarding consumer choices, behaviors, and its outcome (eCommerce360, 2022).

6.3 Model Assumptions and Limitations

Comprehending the assumptions of our preferred model, the logistic regression, is key in achieving the expected outcome. The model assumptions which will aid in the model performance is discussed as follows:

1. Logistic Regression takes only two categorical values, and so the target variable should be binary.
2. The observations of the dataset should be independent of each other, therefore, there should not contain duplicate values. The performance of the model is affected when the features are found to be correlated.
3. There should be none or only a minimal multicollinearity in the data, otherwise it will be a challenge estimating or predicting the variables significant to the target variable.

4. The algorithm is sensitive to outliers and greatly affects the performance of the model. The presence of outliers in the data results to an unexpected outcome, hence should be taken into consideration.
5. Logistic Regression assumes that a linear relationship exists between the independent variables and logit of the target variable.

6.4 Model Sensitivity to Key Drivers

To understand the strong influence on spending habits, the selected variables from the Stepwise Selection method is further discussed in this section.

Selected Variables	Chances of becoming a spender
Per unit increase in Art	Increase by less than .8%
Per unit increase in Physical Activity	Increase by 41%
Per unit increase in Computer Technology	Increase by 8%
Per unit increase in Extrovert	Increase by 2%
Per unit increase in Trendy	Increase by 2.4 times
Per unit increase in Friendliness	Decrease by 3%
Per unit increase in Trustworthiness	Increase by 16%
Per unit increase in Love for animals	Increase by 49%
Per unit increase in Conscientiousness	Decrease by 7%
Per unit increase in Gender (male)	Increase by 28%
Per unit increase in Education (college/bachelor degree)	Decrease by 16%

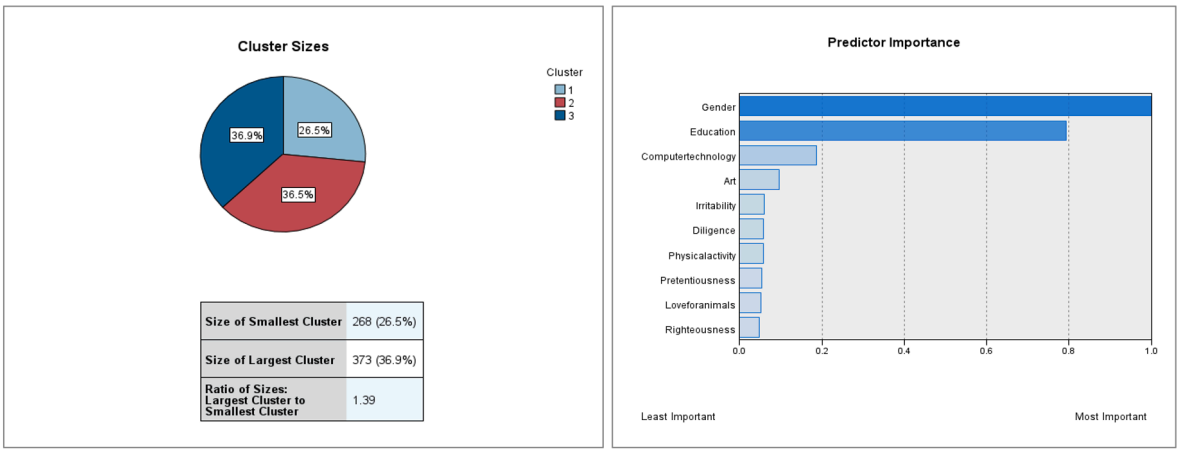
Aside from being trendy, being physically active, and having love for animals, other variables that were found to be significant were art, computer technology,

extrovert, friendliness, trustworthiness, conscientiousness, male gender, and those with college/bachelor's degree. Individuals with art inclinations tend to be on the lower spending possibly due to sophisticated interests. Similarly, people who are "tech-savvy" or well-versed in technology or computers tend to spend higher inorder to stay updated with the latest technology or gadgets. In terms of personality traits, extroverts tend to spend more to enjoy life while trustworthiness, pertaining to those who value quality and reliability, possibly tend to spend more on branded items. On the other hand, individuals who value friends over money tend to spend less and instead devote quality time with their friends. Those who are conscientious, decision-makers or "smart buyers" tend to think twice about their spending and are less likely to spend on material things. Surprisingly, male spends more than women. An article by PwC Slovakia discussed that Slovak women earn 18% less than men (PwC, 2018). Because of this, men potentially spends more than women. Finally, individuals in college or those taking up a bachelor's degree tend to spend less, possibly due to their personal choices or decisions, such as priority over their studies than spending.

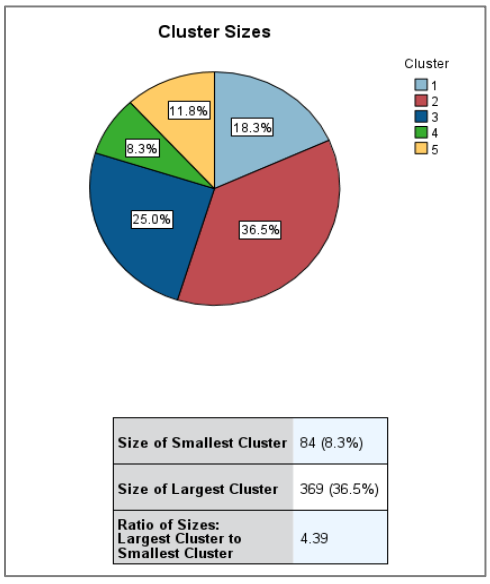
6.5 Cluster Analysis

Aside from identifying significant factors to predict spending, it is also critical to understand certain types of individuals and how ecommerce companies can better reach out and target them. Customer segmentation can be uncovered through cluster analysis. Specifically, a two-step approach in clustering is recommended due to the mixed data types (categorical and continuous) of input variables, which includes demographic variables and personality traits. This approach is available in SPSS; hence, it is utilized for this analysis.

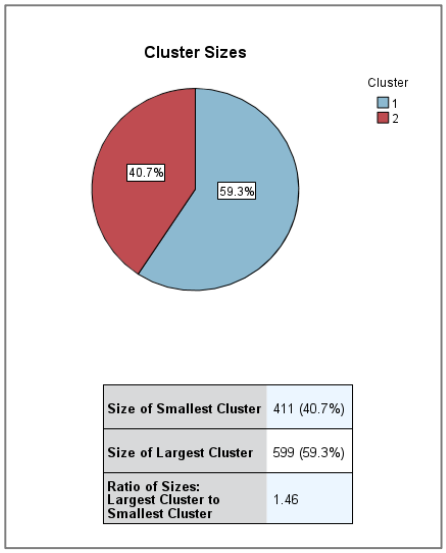
Initial clustering with gender, education, and personality factors as predictors, and using AIC criterion, resulted to three clusters. The categorical variables were the biggest differentiators.



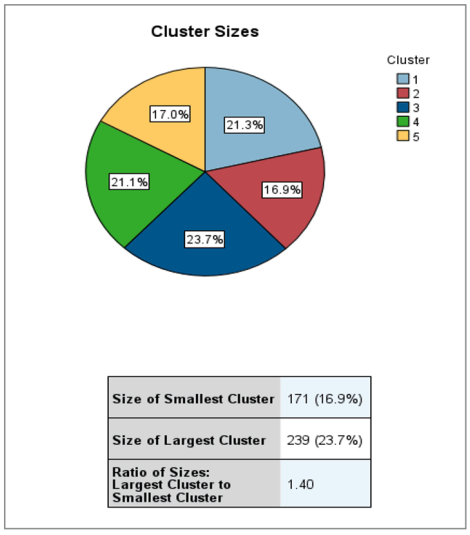
It would be ideal to increase the number of clusters to five to acquire more distinct segments for targeting. However, fixing the number of factors only reflected divisions by gender and education, which is not as beneficial. Rather, the segmentation should be able to capture differences in an individual's behavior.

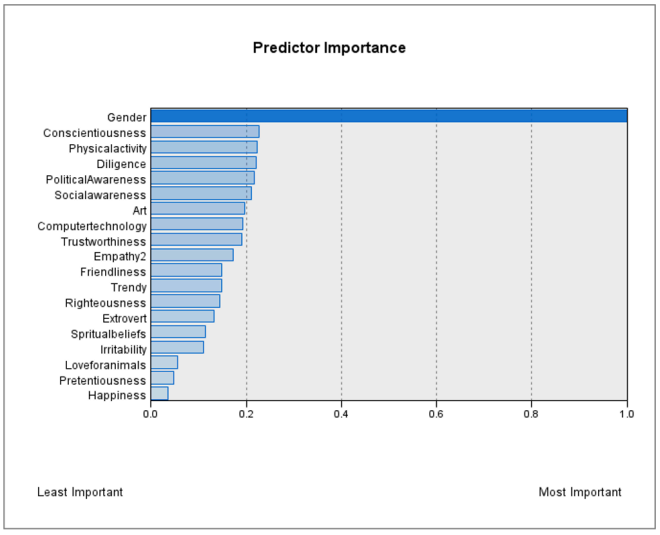


Education was excluded from the predictors to refine the cluster segments, but it only yielded to two cluster segments.



Again, number of factors was further increased to five which seemed to yield better discrimination and varied predictors.





Respondent's belonging to cluster 1 (21%) are females with conscientious and introverted personality. They tend to manage their finances well and thus, are least likely to spend among all the segments.

On the other hand, respondents belonging to cluster 2 (17%), while also females, are more easy-going, are less likely to save, and are moderate spenders. Relative to Gen Z and millennials from other segments, those from this cluster are not as interested in internet and technology devices. However, they are fond of going to shopping malls which implies that they are more likely to make purchases from physical stores than ecommerce.

Cluster 3 (24%) is composed of males who are also less likely to save their money. They are not fond of going to shopping centers, but fond of using the internet and different technology devices. Individuals from this clusters are less concerned about social related matters (law, economy, and management) as well as with being empathetic, being right, nor being spiritual.

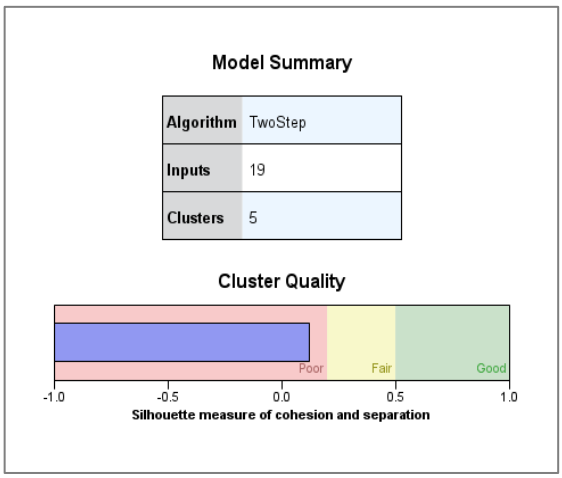
21% of Gen Z and millennials were identified under cluster 4. They are females that are more extroverted, trendy, physically active, and generally more optimistic than other groups. They are more likely to be spenders particularly when it comes to improving their looks and on healthy eating.

Last cluster (17%) includes males with slightly similar profile as females from cluster 4, as they are also found to be extroverted, trendy, and physically active. They are most likely to become heavy spenders among all the segments, exhibiting liking on branded clothing and entertainment spending. They are interested in political and social affairs.

In conclusion, since clusters 4 and 5 are more likely to become high spenders, prioritizing these segments for targeting would be beneficial for ecommerce companies. They take up significant share of Gen Z and millennials at 38% hence targeting them would result to huge profit opportunities. However, it would be valuable to recognize hidden growth opportunities in targeting the rest of Gen Z and millennials

even if they are less spenders since they take up 62% population share. This can be done by planning marketing campaigns and product offerings that would encourage these individuals to increase their purchases. For example, partnership with credit card companies in offering rebates and rewards from ecommerce purchases would appeal to finance-savvy individuals from cluster 1. Web advertisements and retargeting would most likely work for those identified under cluster 3, who are frequent internet and technology users. Gen Z and millennials at cluster 2 are likely to be loyal with physical shopping, hence, most challenging to convert to ecommerce platforms. Partnerships with merchandises in offering promotions and rewards for those who make purchases online could be explored.

It should be noted, however, that due to the nature of most of the Likert-scale questions, above cluster output returned poor cluster quality based on silhouette measure of cohesion and separation indicated below. It is recommended to revisit and validate these clustering groups in future survey repeats. All SPSS syntax for generated cluster analysis is provided in Appendix K.



7.0 VALIDATION AND GOVERNANCE

Data governance is important in all kinds of data analysis in order to achieve high quality of output. Implementing results from these studies would commonly

require huge monetary support and number of resources; hence, it is important to ensure reliability. Rigorous standards and procedures were executed for this study along with thorough cross validation to establish model fit and stability. The summary of governance both in data input and output are outlined in this section.

7.1 Data Input

The data has been thoroughly evaluated to ensure that assumptions are met, and any potential issues are managed before applying any statistical analysis, as follows:

i. **Variable level monitoring:** During the initial data exploration phase, the descriptive statistics was generated to understand the distribution of each continuous variable. First, the range of values were assessed to make sure that everything is within acceptable ranges. Likert questions were all within one to five scale, and all others above five were to be considered as 'Don't Know/ Not Applicable' and treated as missing value. Age values were verified to be aligned with the respondent criteria of fifteen to thirty years old. Height and weight should reflect the same metric, in centimeters and pounds, respectively. Skewness and kurtosis statistics were also included in the descriptive statistics report, and those outside acceptable ranges were flagged for further evaluation. Frequency tables and histograms were also used to check if there's any deviations from normality among the variables.

ii. **Missing rate and Imputation:** Some techniques such as logistic regression is unable to handle missing values in the data. Cases having at least one missing are excluded from the model which result to sample wastage and sample imbalance; hence, it is recommended to employ imputation of missing cases prior to modelling. However, problems would also arise if there were a high

percentage of missing leading to skewed imputed values. A maximum missing rate of 30% was set and any variable that with higher than 30% missing should be excluded from the analysis. For this young adult's survey data, a few variables had small missing percentages of only 1 to 2%, which were imputed using variable mean if numeric or variable mode if categorical.

iii. **Outlier Detection and Treatment (Capping/ Flooring):** Presence of outliers could result to complications in results and compromise its validity. Based on the descriptive statistics table, some of the values were found to have high skewness and kurtosis values which denotes potential outliers. Focus was given on continuous (non-Likert) variables, height and weight. Values above 99^{th} percentile were considered outliers and capped accordingly, while those below 1^{st} percentile were floored.

iv. **Distribution of categories**: Categorical variables in the data should be well-distributed to ensure robustness of each response. 'Doctorate Degree' and 'Master's Degree' education categories were combined to boost response counts, same with 'Currently primary school' and 'Primary school'. *Internetusage* variable was also revised to combine 'No time at all' which only has 3 responses with 'Less than an hour a day'. '*Smoking'* variable was converted to binary, with merged current and former smoker versus those who've never smoked or only tried.

v. **Correlation:** Strong correlation among independent variables can cause problems in the model by introducing incorrect association between predictors and the target variable. Pearson correlation matrix of variables were generated to check if there's any strong association. 'Biology' and 'Medicine' were found

to be moderately correlated at 0.7, while the rest have weak to slightly moderate correlation. All variables passed set benchmark of 0.8.

7.2 Variable Drift Monitoring and Tolerance

Data for this study can be easily reproduced through repeat of the same survey study. However, data characteristics are subject to changes due to several factors. Top sources of changes have been identified in the table below, along with techniques on how these can be monitored and the established model's tolerance level in managing them without drastic loss in reliability and accuracy.

Possible Sources of Data Drift	Monitoring	Tolerance Level
Economic changes resulting to fluctuations in spending habits and preferences	▪ Descriptive Statistics (Mean, Standard Deviation, Range)	***High***: Personality traits are considered universal. Expected reactions to economic fluctuations are gradual, except for certain disruptions.
Changes in data collection methodology	▪ Descriptive Statistics (Mean, Standard Deviation, Range)	***High***: Easily monitored during exploratory data analysis phase. Changes can be remedied such as using cap/ floor technique, standardization, and other transformations.
Changes in demographic distribution of the samples	▪ Cardinality and frequency of demographic variables	***Moderate***: Snowball sampling is used in the survey which might cause differences in sample profile. However, these can be remedied via sample weighting.

7.3 Model Health and Stability

Final model should be easily applied in future datasets and in similar studies, with a comparable level of accuracy. As a first step to ensure that the model is reliable, full dataset were split equally between training and validation. The training dataset was used in building the predictive model in classifying high and low spenders, and accuracy measures are generated to evaluate the model fit. Afterwards, the same model is applied in the validation dataset and the accuracy measure is compared against that of the training dataset. A huge gap in measure between training and validation indicates that the model may be unreliable (overfitting or underfitting). Selected model was found to be stable as it returned a narrow gap between training and validation accuracy, at 62% and 65%, respectively.

Another element affecting model stability is the ease of running the algorithms through the same or even different platforms. The exploratory data analysis was initially run in SPSS and later transferred in Python for ease of replication. Exploratory factor analysis and logistic regression used Python libraries that are easily downloaded and utilized. As for the cluster analysis, the two-step approach used is currently only available in SPSS.

7.4 Model Fit Statistics

The following tests and measures were used to evaluate that, first, appropriate statistical techniques are employed for the dataset, and second, that the model is accurate enough for inference and reproducible. These were decided during the start of the analysis to avoid confirmation bias in selecting optimal model.

i. **KMO and Bartlett's Test:** These tests are valuable in determining whether factor analysis is suited for the data. Young adults dataset returned high KMO value of 0.82, which indicates that there is common information across all

variables that can be translated into factors groupings. This was also supported by a significant p-value output for the Bartlett, which tests whether variables are related or not.

ii. **Average Squared Error (ASE):** Model should yield a low ASE – with a narrow difference between training and validation dataset – to be considered reliable. Selected model has the lowest ASE at 0.35.

iii. **F1-score**: Model should be able to classify cases with high degree of accuracy. This can be determined by the F1-score, which denotes a model's capability of classifying cases as high and low spender correctly. Final model returned average F1-score of 0.60.

8.0 CONCLUSION AND RECOMMENDATION

8.1 Impact on business problems

Findings from this report indicates that personality traits and demographics are good determinants of Gen Z and millennial's spending habits. Major factors affecting whether he/ she would be a low or high spender are *gender, education, interest in art, interest in computer technology, being physically active, being extrovert, being trendy, friendliness, trustworthiness, interest in animals, and conscientiousness.* Being trendy was found to have the strongest association with spend, as those who are more trendy are twice more likely to become high spenders. Interest in sports and animals also increase spending likelihood by at least 40%, while being trustworthy translates to an increased spending likelihood by 16%. The rest of the traits and interests have milder impact on spend by less than 10%, though still significant. It's also remarkable to note that males are 28% more likely to be high spenders than females. Those who have completed college/ bachelor's degree are found to be less likely (16%) to spend, along those who are conscientious and those who would rather invest on time with friends.

While most personality traits are intrinsic, marketing efforts and initiatives are still effective in influencing consumer behavior to yield profit opportunities. For instance, ecommerce companies could consider marketing campaigns that will highlight variety of fashion choices in their website, sending across a message that fashion is for everyone and being trendy does not necessarily need to be expensive. Similarly, ecommerce companies can also aid in promoting the importance of health, consequently encouraging consumers interest in physical activity, and stimulating sales in terms of sports apparel and accessories.

Analyzing and profiling different types of Gen Z and millennials supports findings from the predictive model. Males who are trendy, extroverted, more outgoing, and physically active, were found to have the highest spend rating among all segments. They are particularly interested in spending more on branded clothing and on entertainment. Similarly, females who have similar profile were also found to have high spend rating especially when it comes to improving their looks and healthy eating. These two segments account for a 38% of total Gen Z and millennial population, denoting a profitable opportunity for ecommerce companies if they are to target both segments. Apart from these, it would still be possible and beneficial to reach out to the rest of the segments even if they are less spenders by understanding their behavior, needs, and motivations better. Another segment for male, accounting for 24% of the Gen Z and millennial group, was found to be moderate spenders. They love thriller, sci-fi, and action films, and seems to spend more time on the internet than average; making it easier for ecommerce companies to conduct digital marketing campaigns and retargeting to boost their purchase spending. Another niche segment (17%) composed of females were found to be less savvy in dealing in managing finances. Though not as trendy, they still exhibit interest in spending for their looks. They are not

as interested in computer technology and still prefers going to shopping malls; hence, they are more likely to buy from physical stores rather than online. Ecommerce companies could explore strategies on how to convert these females to purchase online. Companies could try to find ways on improving their web interfaces to make it more attractive and friendly to those who are not fond of online purchases, as well as offering product bundles and discounts that are not available in-stores. The last segment composed of also females (21%) could be the most challenging to penetrate. They were found to be introverted, conscientious, and less likely to make impulse decisions. They are also the most finance-savvy among all the segments. Emphasizing discounts and rebates would be the most effective strategy for these Gen Z and millennials.

Conjoining the results from the segmentation and from the predictive model would be a powerful tool for companies in achieving its business objective on increasing penetration, market share, and profits. The predictive model serves as an aid in identifying the factors that would influence Gen Z and millennial's spending habit and classifying whether one is high or low spender. Whereas the identified segments can be used in customizing marketing campaigns according to customer behavior and profile.

8.2 Recommendations

Ecommerce companies are recommended to pursue the following actions:

1. For future studies, it is advisable to run the logistic regression method to each of the identified cluster groups. Doing so will aid in verifying if there are any differences in the predictors that significantly influence Gen Z and millennials' spending behaviour across segments.

2. Conducting quantitative and qualitative research studies would be beneficial to better understand goals and motivations of Gen Z and millennials. These studies would provide in-depth insight on how certain predictors could be utilized to help young adults achieve their objectives. For example, health-related aspirations could be associated with improved physical activity.
3. Conceptualize loyalty programs for identified segment of Gen Z and millennials who are more likely to become high spenders. As these groups tend to be the "best" or the "most valuable" customers since they generate far better profit compared to the rest, ecommerce companies should find ways to maintain customer loyalty. In order to do that, it is recommended for ecommerce companies to show their appreciation through personalized communication and offers, including continuous innovation on ways that will make their shopping experience always pleasant. Doing so would result to having total customer experience and huge profit opportunities for the business.
4. Create customized marketing campaigns and personalized product offerings, along with user interface A/B testing, to encourage segments that are hesitant to convert to online purchase as well as those with low purchase spending.
5. Establish partnerships with merchants and financial institutions to build efforts in offering the best rewards for consumers who purchase online.

APPENDICES

Appendix A

SPSS syntax used for initial data investigation

```
* Encoding: UTF-8.

/*EDA for Capstone (Manzano and Rimando)

/*Importing excel dataset from kaggle

GET DATA
  /TYPE=XLSX
  /FILE='C:\Users\wmanzano\Documents\Capstone\Young Adults Survey Data
- Capstone.xlsx'
  /SHEET=name 'responses'
  /CELLRANGE=FULL
  /READNAMES=ON
  /DATATYPEMIN PERCENTAGE=95.0
  /HIDDEN IGNORE=YES.
EXECUTE.
DATASET NAME DataSet1 WINDOW=FRONT.
a

SAVE OUTFILE='C:\Users\wmanzano\Documents\Capstone\Capstone - Young
Adults Data.sav'
  /COMPRESSED.

/*Converting string variables to categorical - numeric

DATASET ACTIVATE DataSet1.
AUTORECODE VARIABLES=Smoking Alcohol Punctuality Lying Gender
Leftrighthanded Education Onlychild
    Villagetown Houseblockofflats
  /INTO smoking_recode alcohol_recode punctual_recode lying_recode
gender_recode hand_recode
    education_recode onlychild_recode villagetown_recode house_recode
  /BLANK=MISSING
  /PRINT.

/*Adding unique ID

COMPUTE ID=$CASENUM.
EXECUTE.

/*Running descriptive statistics to check variable range and
distribution

DESCRIPTIVES VARIABLES=Music Slowsongsorfastsongs Dance Folk Country
Classicalmusic Musical Pop
    Rock MetalorHardrock Punk HiphopRap ReggaeSka SwingJazz Rocknroll
Alternative Latino TechnoTrance
    Opera Movies Horror Thriller Comedy Romantic Scifi War
FantasyFairytales Animated Documentary
```

```
    Western Action History Psychology Politics Mathematics Physics
Internet PC EconomyManagement
    Biology Chemistry Reading Geography Foreignlanguages Medicine Law
Cars Artexhibitions Religion
    Countrysideoutdoors Dancing Musicalinstruments Writing Passivesport
Activesport Gardening
    Celebrities Shopping Scienceandtechnology Theatre Funwithfriends
Adrenalinesports Pets Flying Storm
    Darkness Heights Spiders Snakes Rats Ageing Dangerousdogs
Fearofpublicspeaking Healthyeating
    Dailyevents Prioritisingworkload Writingnotes Workaholism
Thinkingahead Finaljudgement Reliability
    Keepingpromises Lossofinterest Friendsversusmoney Funniness Fake
Criminaldamage Decisionmaking
    Elections Selfcriticism Judgmentcalls Hypochondria Empathy
Eatingtosurvive Giving
    Compassiontoanimals Borrowedstuff Loneliness Cheatinginschool
Health Changingthepast God Dreams
    Charity Numberoffriends Waiting Newenvironment Moodswings
Appearenceandgestures Socializing
    Achievements Respondingtoaseriousletter Children Assertiveness
Gettingangry Knowingtherightpeople
    Publicspeaking Unpopularity Lifestruggles Happinessinlife
Energylevels Smallbigdogs Personality
    Findinglostvaluables Gettingup Interestsorhobbies Parentsadvice
Questionnairesorpolls Finances
    Shoppingcentres Brandedclothing Entertainmentspending
Spendingonlooks Spendingongadgets
    Spendingonhealthyeating Age Height Weight Numberofsiblings
smoking_recode alcohol_recode
    punctual_recode lying_recode gender_recode hand_recode
education_recode onlychild_recode
    villagetown_recode house_recode ID
  /STATISTICS=MEAN STDDEV MIN MAX KURTOSIS SKEWNESS.

/*Generating frequency tables and histogram for all variables to check
distribution

FREQUENCIES VARIABLES=Music Slowsongsorfastsongs Dance Folk Country
Classicalmusic Musical Pop Rock
    MetalorHardrock Punk HiphopRap ReggaeSka SwingJazz Rocknroll
Alternative Latino TechnoTrance Opera
    Movies Horror Thriller Comedy Romantic Scifi War FantasyFairytales
Animated Documentary Western
    Action History Psychology Politics Mathematics Physics Internet PC
EconomyManagement Biology
    Chemistry Reading Geography Foreignlanguages Medicine Law Cars
Artexhibitions Religion
    Countrysideoutdoors Dancing Musicalinstruments Writing Passivesport
Activesport Gardening
    Celebrities Shopping Scienceandtechnology Theatre Funwithfriends
Adrenalinesports Pets Flying Storm
    Darkness Heights Spiders Snakes Rats Ageing Dangerousdogs
Fearofpublicspeaking Smoking Alcohol
    Healthyeating Dailyevents Prioritisingworkload Writingnotes
Workaholism Thinkingahead
    Finaljudgement Reliability Keepingpromises Lossofinterest
Friendsversusmoney Funniness Fake
```

```
    Criminaldamage Decisionmaking Elections Selfcriticism Judgmentcalls
Hypochondria Empathy
    Eatingtosurvive Giving Compassiontoanimals Borrowedstuff Loneliness
Cheatinginschool Health
    Changingthepast God Dreams Charity Numberoffriends Punctuality
Lying Waiting Newenvironment
    Moodswings Appearenceandgestures Socializing Achievements
Respondingtoaseriousletter Children
    Assertiveness Gettingangry Knowingtherightpeople Publicspeaking
Unpopularity Lifestruggles
    Happinessinlife Energylevels Smallbigdogs Personality
Findinglostvaluables Gettingup
    Interestsorhobbies Parentsadvice Questionnairesorpolls
Internetusage Finances Shoppingcentres
    Brandedclothing Entertainmentspending Spendingonlooks
Spendingongadgets Spendingonhealthyeating Age
    Height Weight Numberofsiblings Gender Leftrighthanded Education
Onlychild Villagetown
    Houseblockofflats smoking_recode alcohol_recode punctual_recode
lying_recode gender_recode
    hand_recode education_recode onlychild_recode villagetown_recode
house_recode ID
  /HISTOGRAM NORMAL
  /ORDER=ANALYSIS.

/*Checking high correlation

CORRELATIONS
  /VARIABLES=Music Slowsongsorfastsongs Dance Folk Country
Classicalmusic Musical Pop Rock
    MetalorHardrock Punk HiphopRap ReggaeSka SwingJazz Rocknroll
Alternative Latino TechnoTrance Opera
    Movies Horror Thriller Comedy Romantic Scifi War FantasyFairytales
Animated Documentary Western
    Action History Psychology Politics Mathematics Physics Internet PC
EconomyManagement Biology
    Chemistry Reading Geography Foreignlanguages Medicine Law Cars
Artexhibitions Religion
    Countrysideoutdoors Dancing Musicalinstruments Writing Passivesport
Activesport Gardening
    Celebrities Shopping Scienceandtechnology Theatre Funwithfriends
Adrenalinesports Pets Flying Storm
    Darkness Heights Spiders Snakes Rats Ageing Dangerousdogs
Fearofpublicspeaking Healthyeating
    Dailyevents Prioritisingworkload Writingnotes Workaholism
Thinkingahead Finaljudgement Reliability
    Keepingpromises Lossofinterest Friendsversusmoney Funniness Fake
Criminaldamage Decisionmaking
    Elections Selfcriticism Judgmentcalls Hypochondria Empathy
Eatingtosurvive Giving
    Compassiontoanimals Borrowedstuff Loneliness Cheatinginschool
Health Changingthepast God Dreams
    Charity Numberoffriends Waiting Newenvironment Moodswings
Appearenceandgestures Socializing
    Achievements Respondingtoaseriousletter Children Assertiveness
Gettingangry Knowingtherightpeople
    Publicspeaking Unpopularity Lifestruggles Happinessinlife
Energylevels Smallbigdogs Personality
```

```
    Findinglostvaluables Gettingup Interestsorhobbies Parentsadvice
Questionnairesorpolls Finances
    Shoppingcentres Brandedclothing Entertainmentspending
Spendingonlooks Spendingongadgets
    Spendingonhealthyeating Age Height Weight Numberofsiblings
smoking_recode alcohol_recode
    punctual_recode lying_recode gender_recode hand_recode
education_recode onlychild_recode
    villagetown_recode house_recode ID
  /PRINT=TWOTAIL SIG FULL
  /MISSING=PAIRWISE.

/*Creating target variable (mean spending rating)

COMPUTE
MeanSpendRating=MEAN(Shoppingcentres,Brandedclothing,Entertainmentspend
ing,Spendingonlooks,
    Spendingongadgets,Spendingonhealthyeating).
EXECUTE.

/*checking frequency

FREQUENCIES VARIABLES=MeanSpendRating
  /HISTOGRAM NORMAL
  /ORDER=ANALYSIS.

/*T-test mean spend rating between gender

T-TEST GROUPS=gender_recode(1 2)
  /MISSING=ANALYSIS
  /VARIABLES=MeanSpendRating
  /ES DISPLAY(FALSE)
  /CRITERIA=CI(.95).

/* Generating Dummy Variables for Regression

SPSSINC CREATE DUMMIES VARIABLE=smoking_recode alcohol_recode
punctual_recode lying_recode
    gender_recode hand_recode education_recode onlychild_recode
villagetown_recode house_recode
ROOTNAME1=smoking alcohol punctual lying gender hand educ onlychild
town house
/OPTIONS ORDER=A USEVALUELABELS=YES USEML=NO OMITFIRST=YES.

/*Generating initial model (all variables excluding individual spend
variables, phobias, and left hand/ right hand, shopping)

REGRESSION
  /MISSING MEANSUBSTITUTION
  /STATISTICS COEFF OUTS R ANOVA COLLIN TOL
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT MeanSpendRating
  /METHOD=STEPWISE Music Slowsongsorfastsongs Dance Folk Country
Classicalmusic Musical Pop Rock
    MetalorHardrock Punk HiphopRap ReggaeSka SwingJazz Rocknroll
Alternative Latino TechnoTrance Opera
```

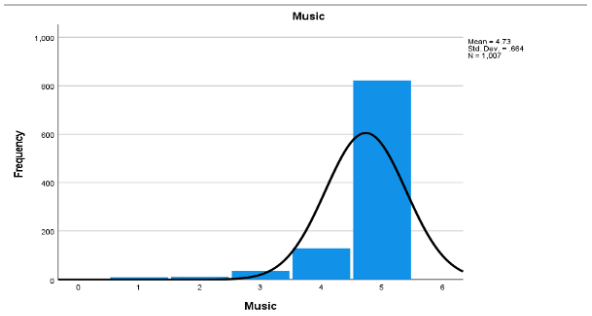
```
    Movies Horror Thriller Comedy Romantic Scifi War FantasyFairytales
Animated Documentary Western
    Action History Psychology Politics Mathematics Physics Internet PC
EconomyManagement Biology
    Chemistry Reading Geography Foreignlanguages Medicine Law Cars
Artexhibitions Religion
    Countrysideoutdoors Dancing Musicalinstruments Writing Passivesport
Activesport Gardening
    Celebrities Scienceandtechnology Theatre Funwithfriends
Adrenalinesports Pets Healthyeating
    Dailyevents Prioritisingworkload Writingnotes Workaholism
Thinkingahead Finaljudgement Reliability
    Keepingpromises Lossofinterest Friendsversusmoney Funniness Fake
Criminaldamage Decisionmaking
    Elections Selfcriticism Judgmentcalls Hypochondria Empathy
Eatingtosurvive Giving
    Compassiontoanimals Borrowedstuff Loneliness Cheatinginschool
Health Changingthepast God Dreams
    Charity Numberoffriends Waiting Newenvironment Moodswings
Appearenceandgestures Socializing
    Achievements Respondingtoaseriousletter Children Assertiveness
Gettingangry Knowingtherightpeople
    Publicspeaking Unpopularity Lifestruggles Happinessinlife
Energylevels Smallbigdogs Personality
    Findinglostvaluables Gettingup Interestsorhobbies Parentsadvice
Questionnairesorpolls Finances Age
    Height Weight Numberofsiblings smoking_1 smoking_2 smoking_3
alcohol_6 alcohol_7 punctual_10
    punctual_11 lying_14 lying_15 lying_16 gender_19 hand_22 educ_25
educ_26 educ_27 educ_28 educ_29
    town_35 house_38.

/*Generating initial factor analysis

/*VARIMAX ROTATION

FACTOR
  /VARIABLES Dailyevents Prioritisingworkload Writingnotes Workaholism
Thinkingahead Finaljudgement
    Reliability Keepingpromises Lossofinterest Friendsversusmoney
Funniness Fake Criminaldamage
    Decisionmaking Elections Selfcriticism Judgmentcalls Hypochondria
Empathy Eatingtosurvive Giving
    Compassiontoanimals Borrowedstuff Loneliness Cheatinginschool
Health Changingthepast God Dreams
    Charity Numberoffriends Waiting Newenvironment Moodswings
Appearenceandgestures Socializing
    Achievements Respondingtoaseriousletter Children Assertiveness
Gettingangry Knowingtherightpeople
    Publicspeaking Unpopularity Lifestruggles Happinessinlife
Energylevels Smallbigdogs Personality
    Findinglostvaluables Gettingup Interestsorhobbies Parentsadvice
Questionnairesorpolls
  /MISSING MEANSUB
  /ANALYSIS Dailyevents Prioritisingworkload Writingnotes Workaholism
Thinkingahead Finaljudgement
    Reliability Keepingpromises Lossofinterest Friendsversusmoney
Funniness Fake Criminaldamage
```

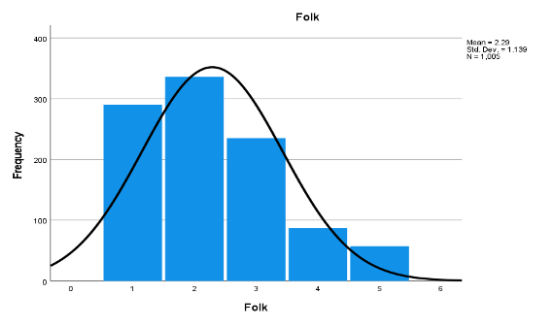
```
    Decisionmaking Elections Selfcriticism Judgmentcalls Hypochondria
Empathy Eatingtosurvive Giving
    Compassiontoanimals Borrowedstuff Loneliness Cheatinginschool
Health Changingthepast God Dreams
    Charity Numberoffriends Waiting Newenvironment Moodswings
Appearenceandgestures Socializing
    Achievements Respondingtoaseriousletter Children Assertiveness
Gettingangry Knowingtherightpeople
    Publicspeaking Unpopularity Lifestruggles Happinessinlife
Energylevels Smallbigdogs Personality
    Findinglostvaluables Gettingup Interestsorhobbies Parentsadvice
Questionnairesorpolls
  /PRINT INITIAL KMO EXTRACTION ROTATION
  /FORMAT BLANK(.2)
  /PLOT EIGEN
  /CRITERIA MINEIGEN(1) ITERATE(25)
  /EXTRACTION PAF
  /CRITERIA ITERATE(25)
  /ROTATION VARIMAX
  /METHOD=CORRELATION.

/*EQUAMAX

FACTOR
  /VARIABLES Dailyevents Prioritisingworkload Writingnotes Workaholism
Thinkingahead Finaljudgement
    Reliability Keepingpromises Lossofinterest Friendsversusmoney
Funniness Fake Criminaldamage
    Decisionmaking Elections Selfcriticism Judgmentcalls Hypochondria
Empathy Eatingtosurvive Giving
    Compassiontoanimals Borrowedstuff Loneliness Cheatinginschool
Health Changingthepast God Dreams
    Charity Numberoffriends Waiting Newenvironment Moodswings
Appearenceandgestures Socializing
    Achievements Respondingtoaseriousletter Children Assertiveness
Gettingangry Knowingtherightpeople
    Publicspeaking Unpopularity Lifestruggles Happinessinlife
Energylevels Smallbigdogs Personality
    Findinglostvaluables Gettingup Interestsorhobbies Parentsadvice
Questionnairesorpolls
  /MISSING MEANSUB
  /ANALYSIS Dailyevents Prioritisingworkload Writingnotes Workaholism
Thinkingahead Finaljudgement
    Reliability Keepingpromises Lossofinterest Friendsversusmoney
Funniness Fake Criminaldamage
    Decisionmaking Elections Selfcriticism Judgmentcalls Hypochondria
Empathy Eatingtosurvive Giving
    Compassiontoanimals Borrowedstuff Loneliness Cheatinginschool
Health Changingthepast God Dreams
    Charity Numberoffriends Waiting Newenvironment Moodswings
Appearenceandgestures Socializing
    Achievements Respondingtoaseriousletter Children Assertiveness
Gettingangry Knowingtherightpeople
    Publicspeaking Unpopularity Lifestruggles Happinessinlife
Energylevels Smallbigdogs Personality
    Findinglostvaluables Gettingup Interestsorhobbies Parentsadvice
Questionnairesorpolls
  /PRINT INITIAL KMO EXTRACTION ROTATION
```

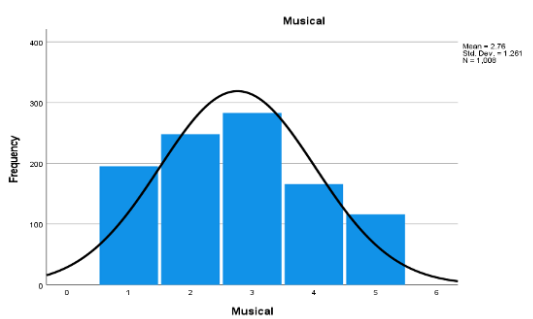
```
  /FORMAT BLANK(.2)
  /PLOT EIGEN
  /CRITERIA MINEIGEN(1) ITERATE(25)
  /EXTRACTION PAF
  /CRITERIA ITERATE(25)
  /ROTATION EQUAMAX
  /METHOD=CORRELATION.

/*Creating binary spend variable (spender = 4, 5 mean spend; non-
spender = 1, 2 or 3)

RECODE MeanSpendRating (4 thru 5=1) (ELSE=0) INTO Spend_Binary.
VARIABLE LABELS Spend_Binary 'Spend_Binary'.
EXECUTE.
```

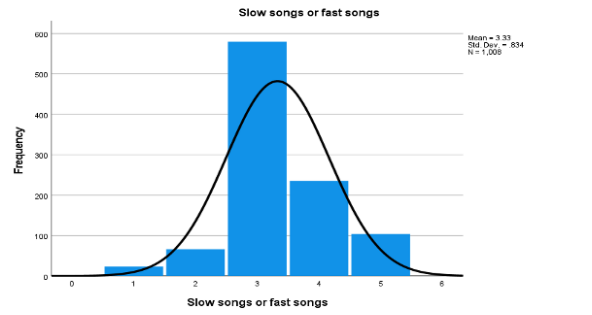
Appendix B

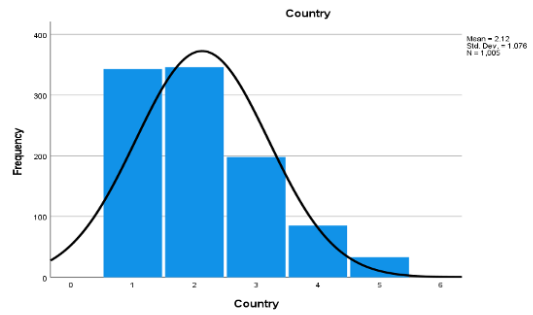
Histogram of all the variables in the Young Adult dataset

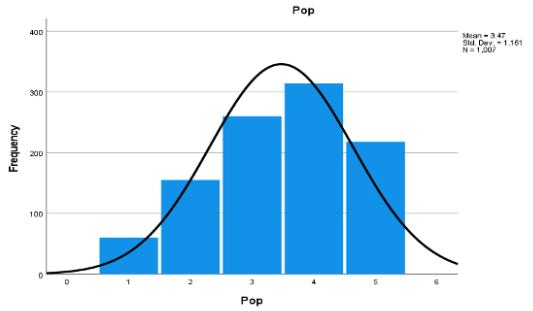


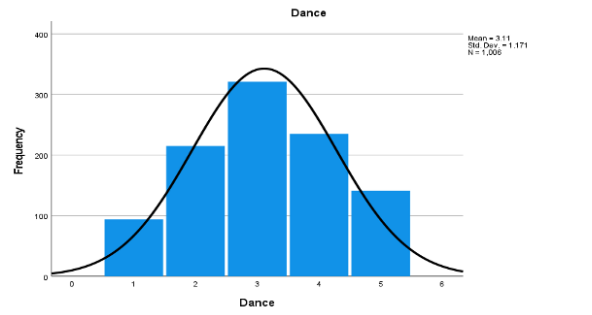


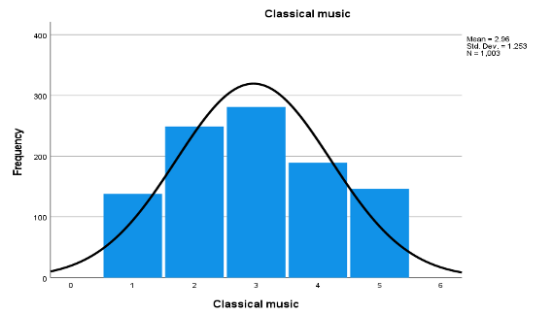


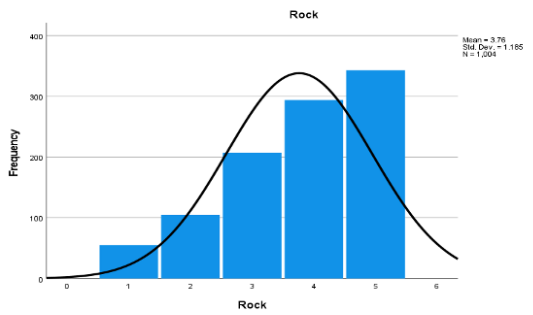


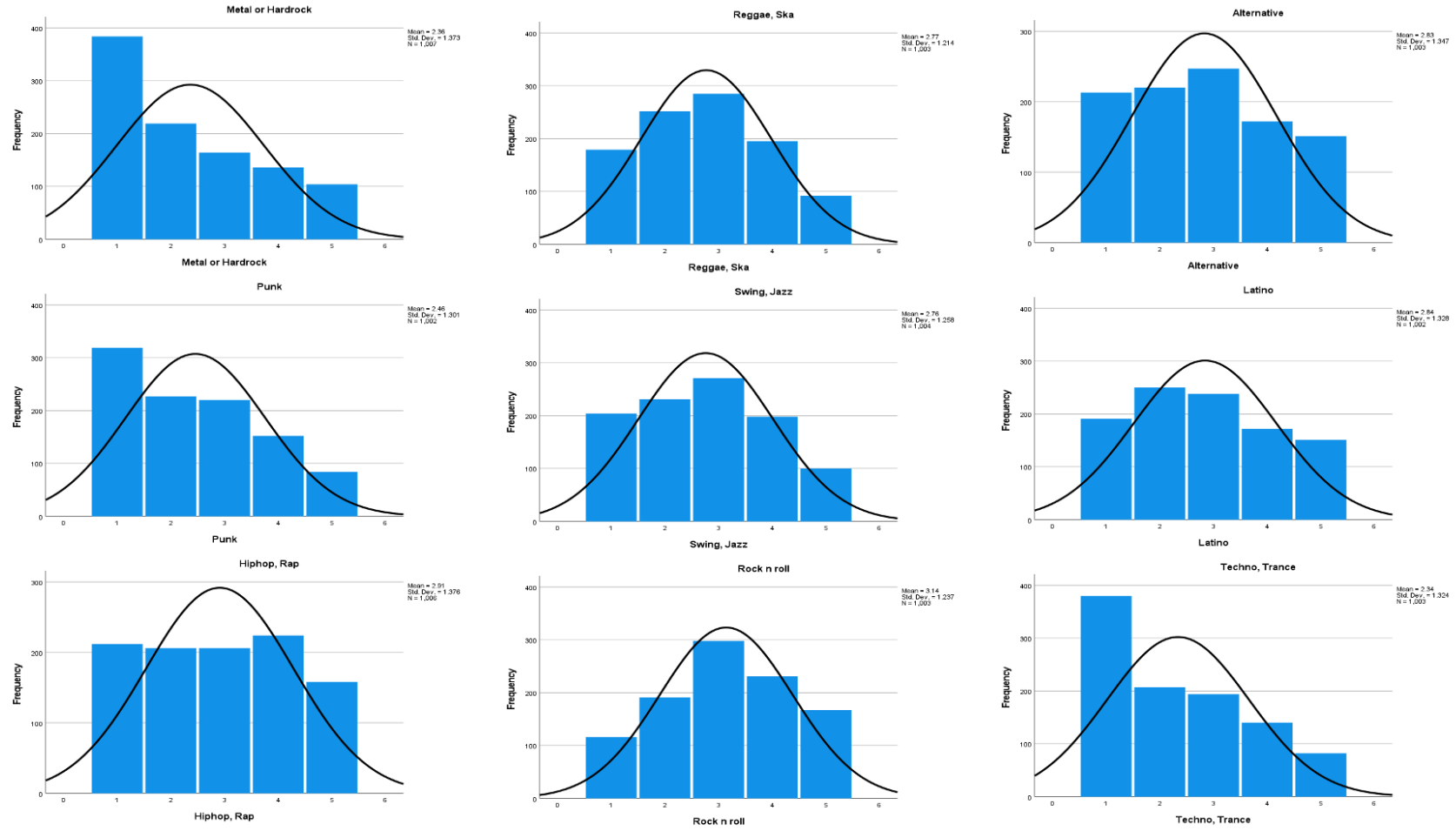


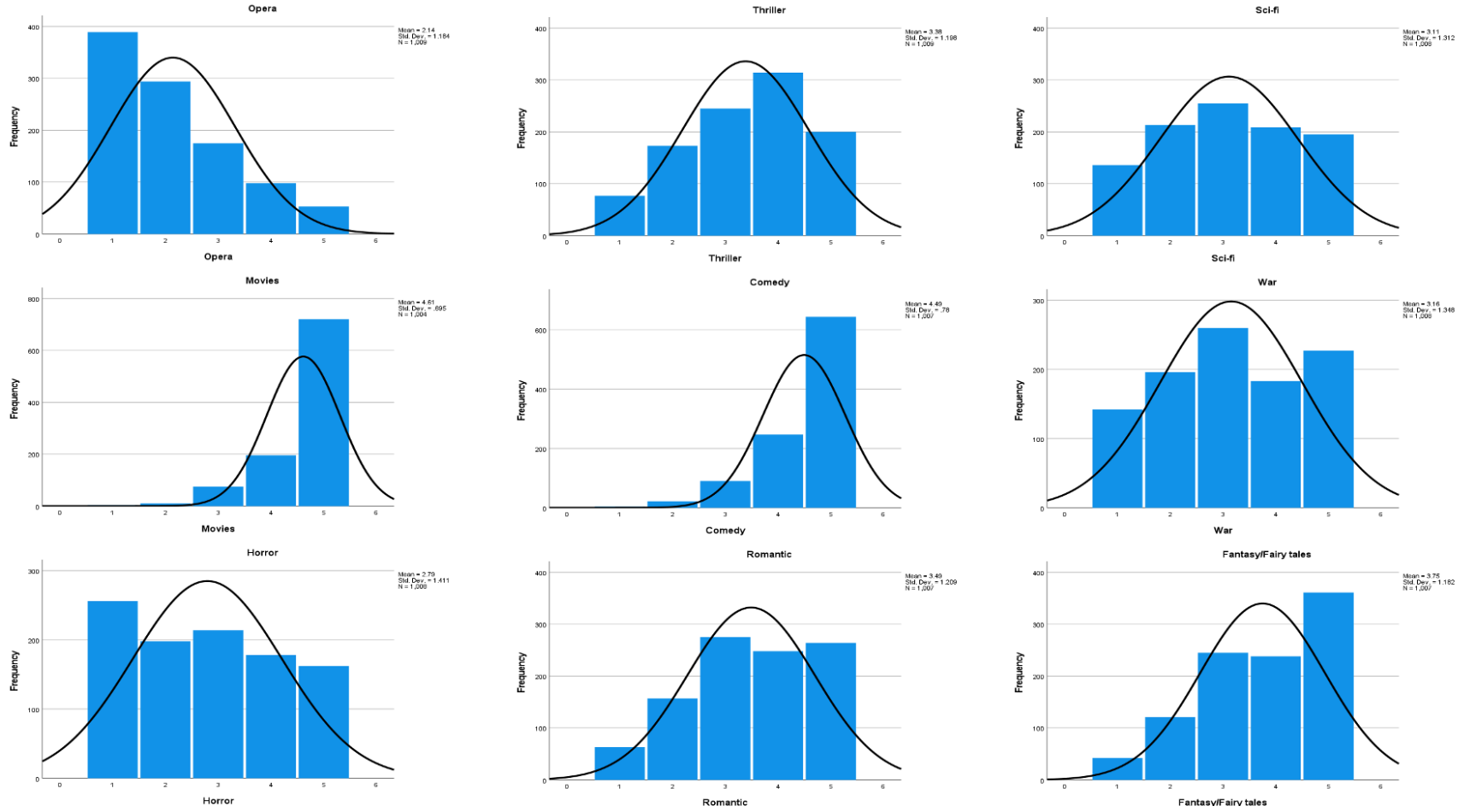


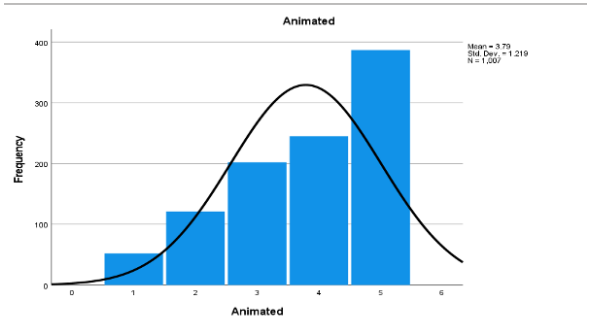


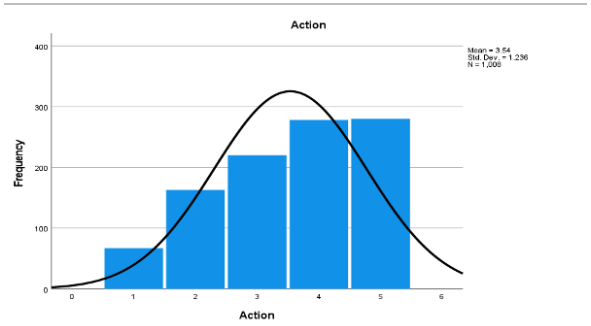


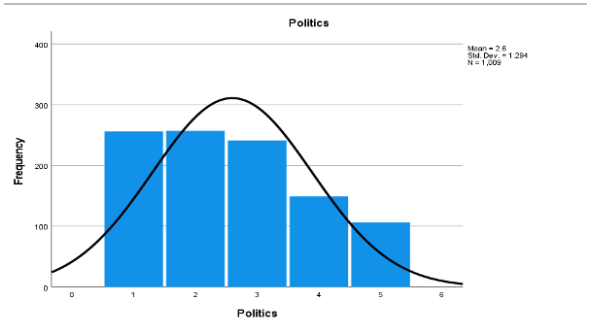


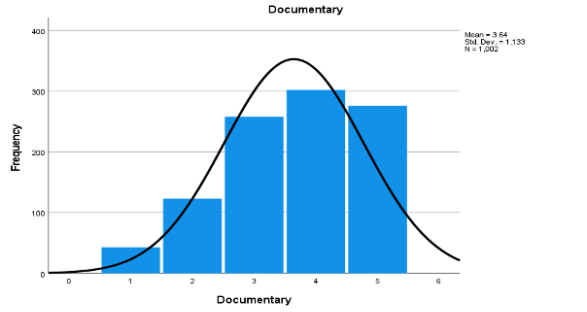

Metal or Hardrock
Frequency
Metal or Hardrock
Reggae, Ska
Frequency
Reggae, Ska
Alternative
Frequency
Alternative
Punk
Frequency
Punk
Swing, Jazz
Frequency
Swing, Jazz
Latino
Frequency
Latino
Hiphop, Rap
Frequency
Hiphop, Rap
Rock n roll
Frequency
Rock n roll
Techno, Trance
Frequency
Techno, Trance

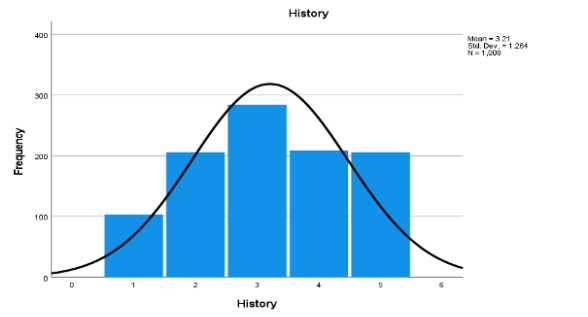

Opera
Frequency
Opera
Thriller
Frequency
Thriller
Sci-fi
Frequency
Sci-fi
Movies
Frequency
Movies
Comedy
Frequency
Comedy
War
Frequency
War
Horror
Frequency
Horror
Romantic
Frequency
Romantic
Fantasy/Fairy tales
Frequency
Fantasy/Fairy tales

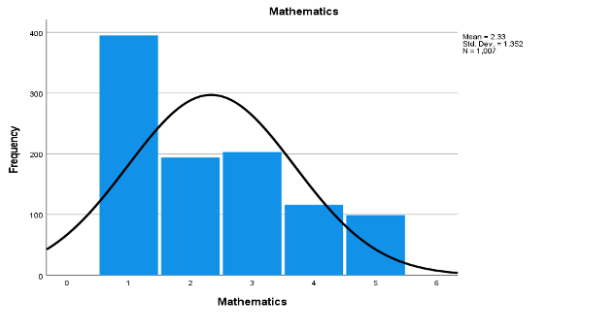

Animated
Frequency
Animated

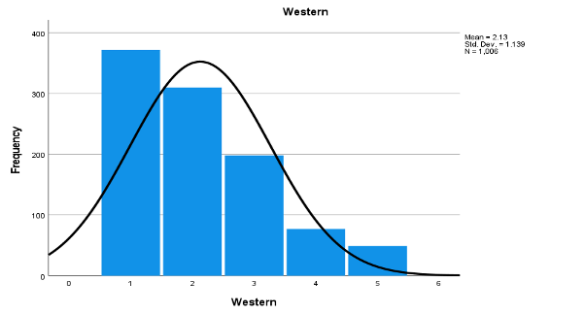

Action
Frequency
Action

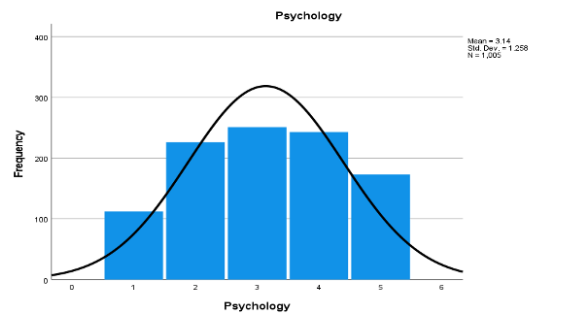

Politics
Frequency
Politics

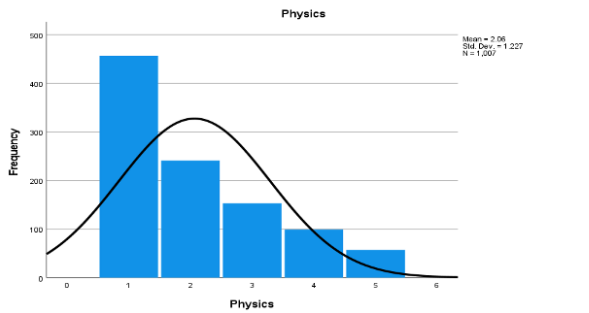

Documentary
Frequency
Documentary

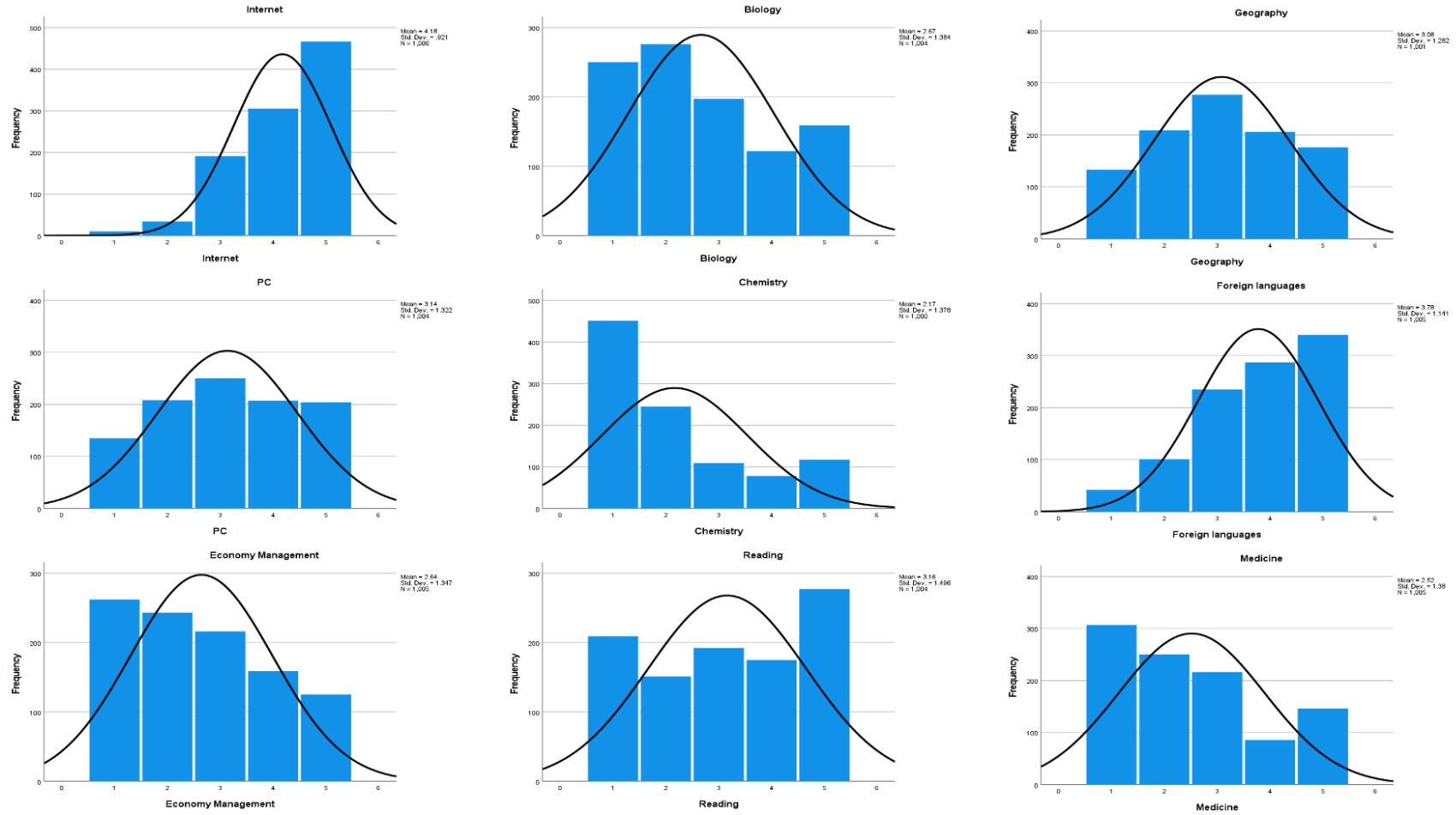

History
Frequency
History

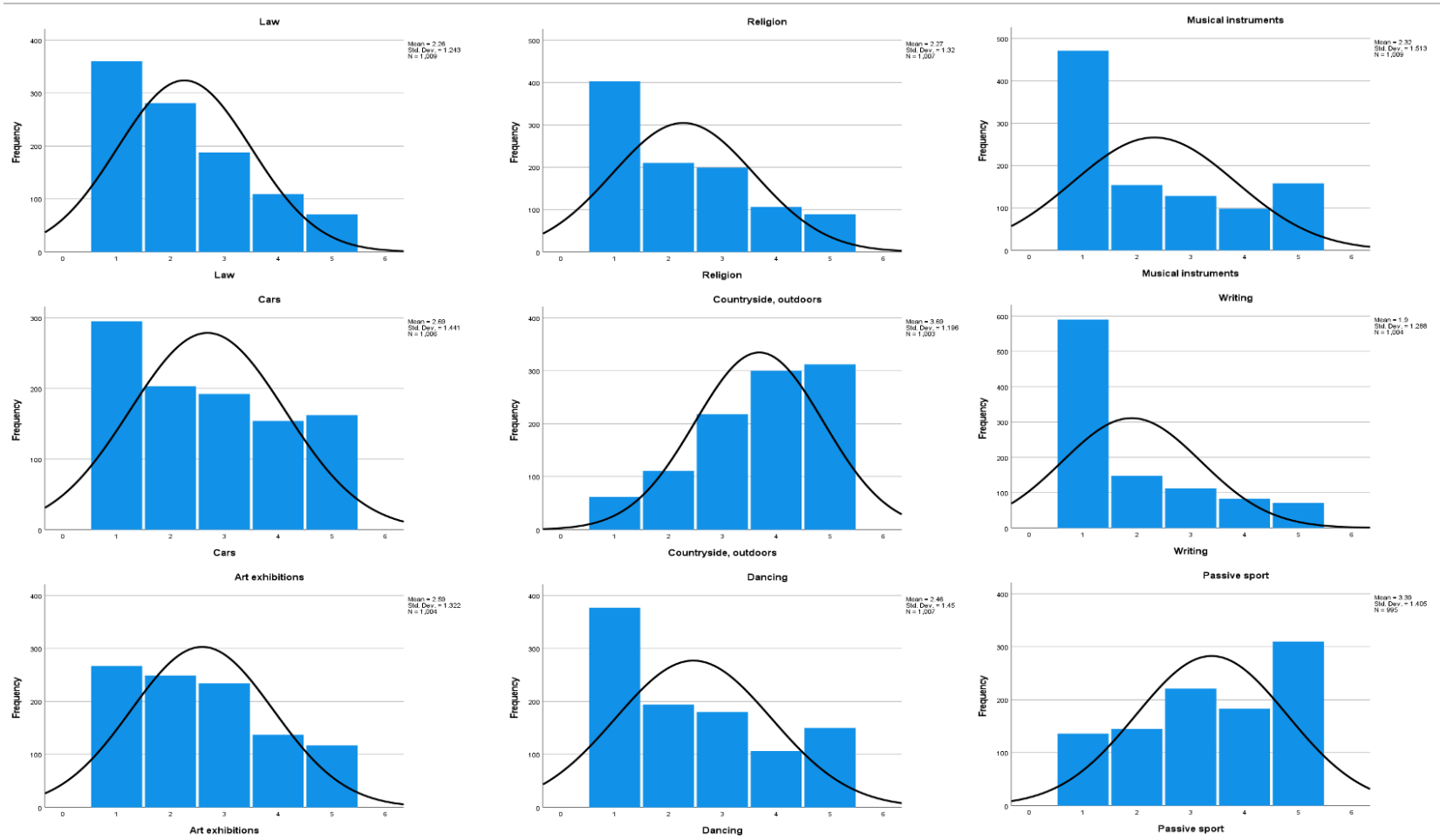

Mathematics
Frequency
Mathematics

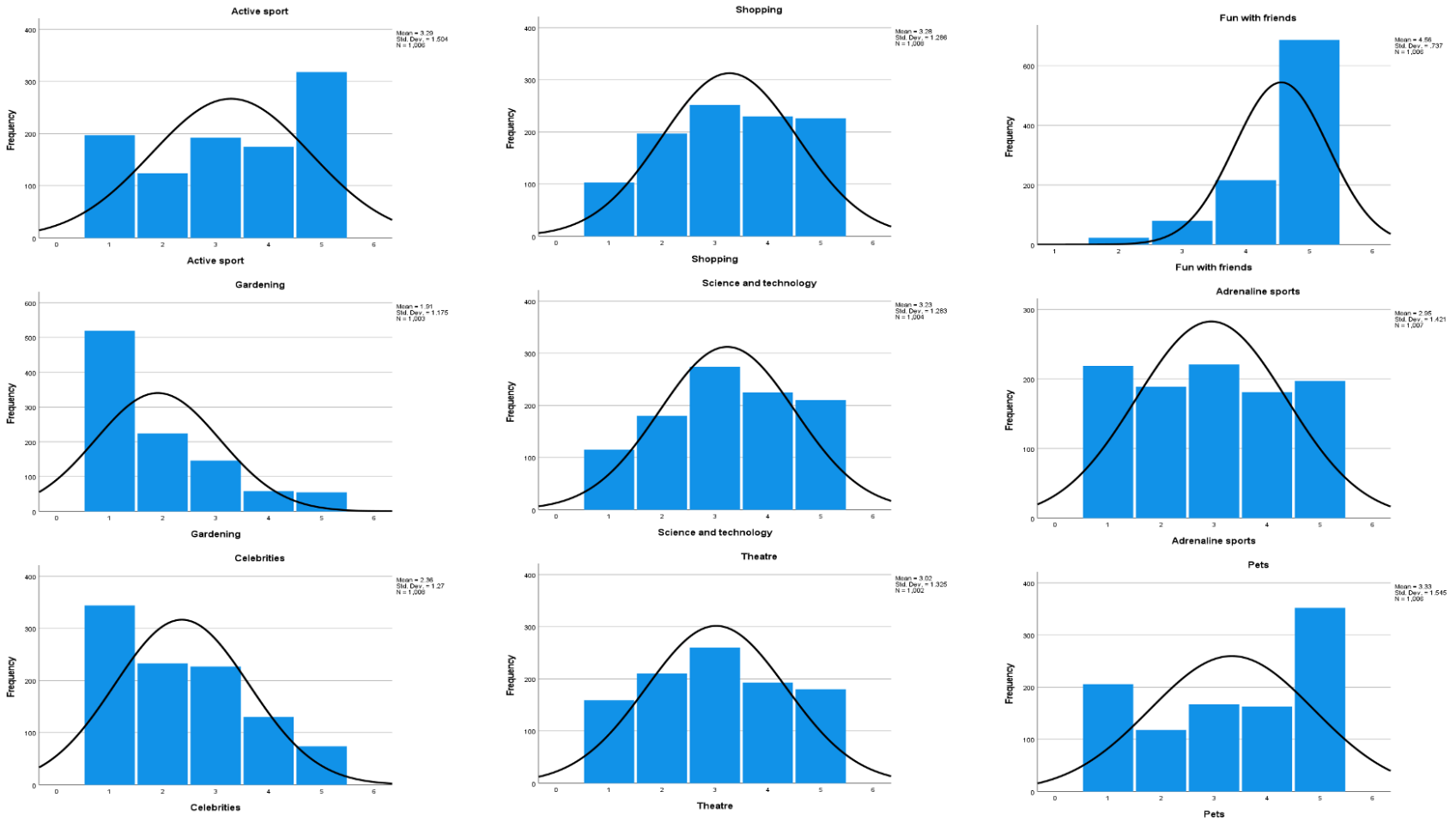

Western
Frequency
Western

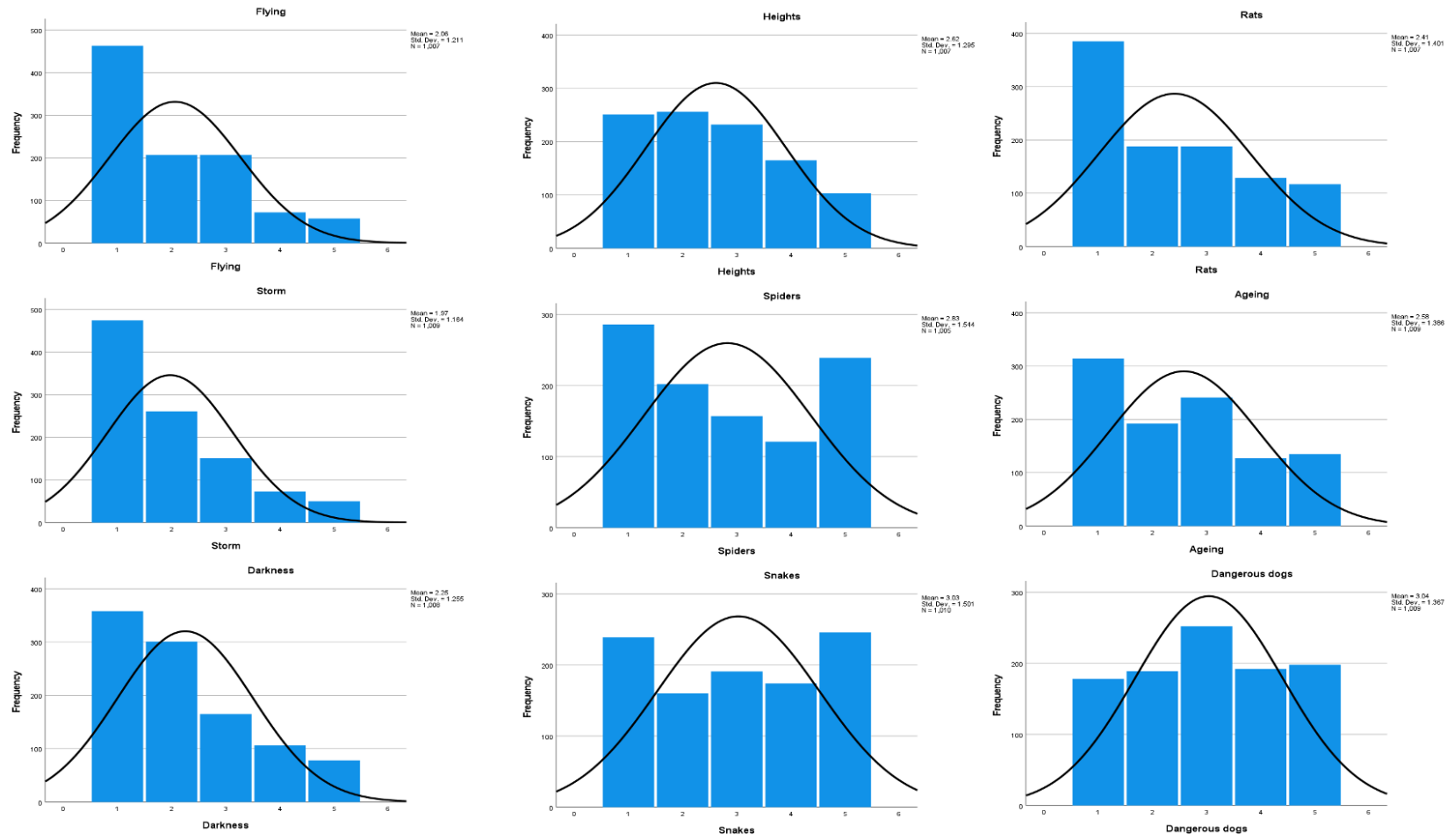

Psychology
Frequency
Psychology

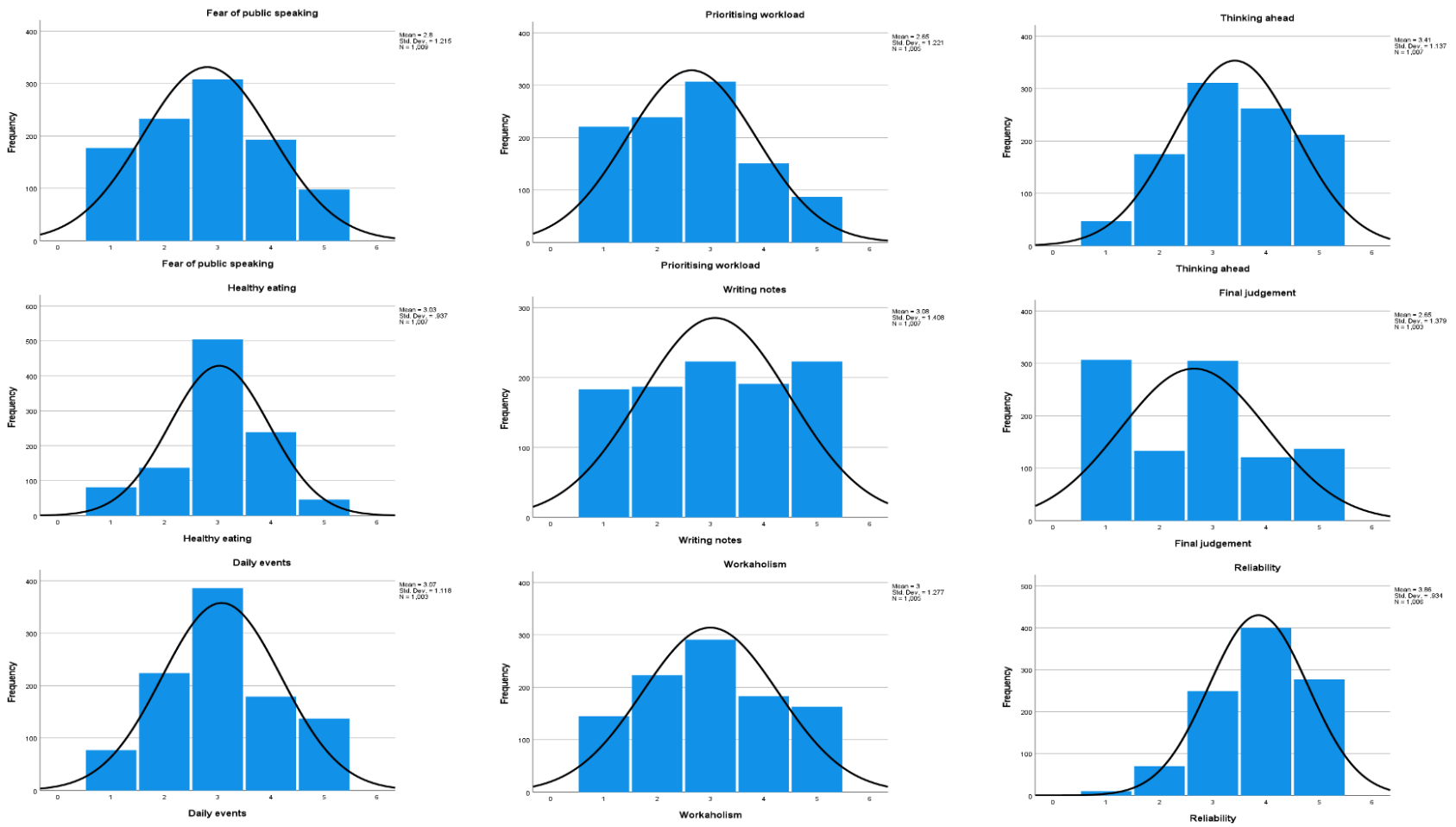

Physics
Frequency
Physics

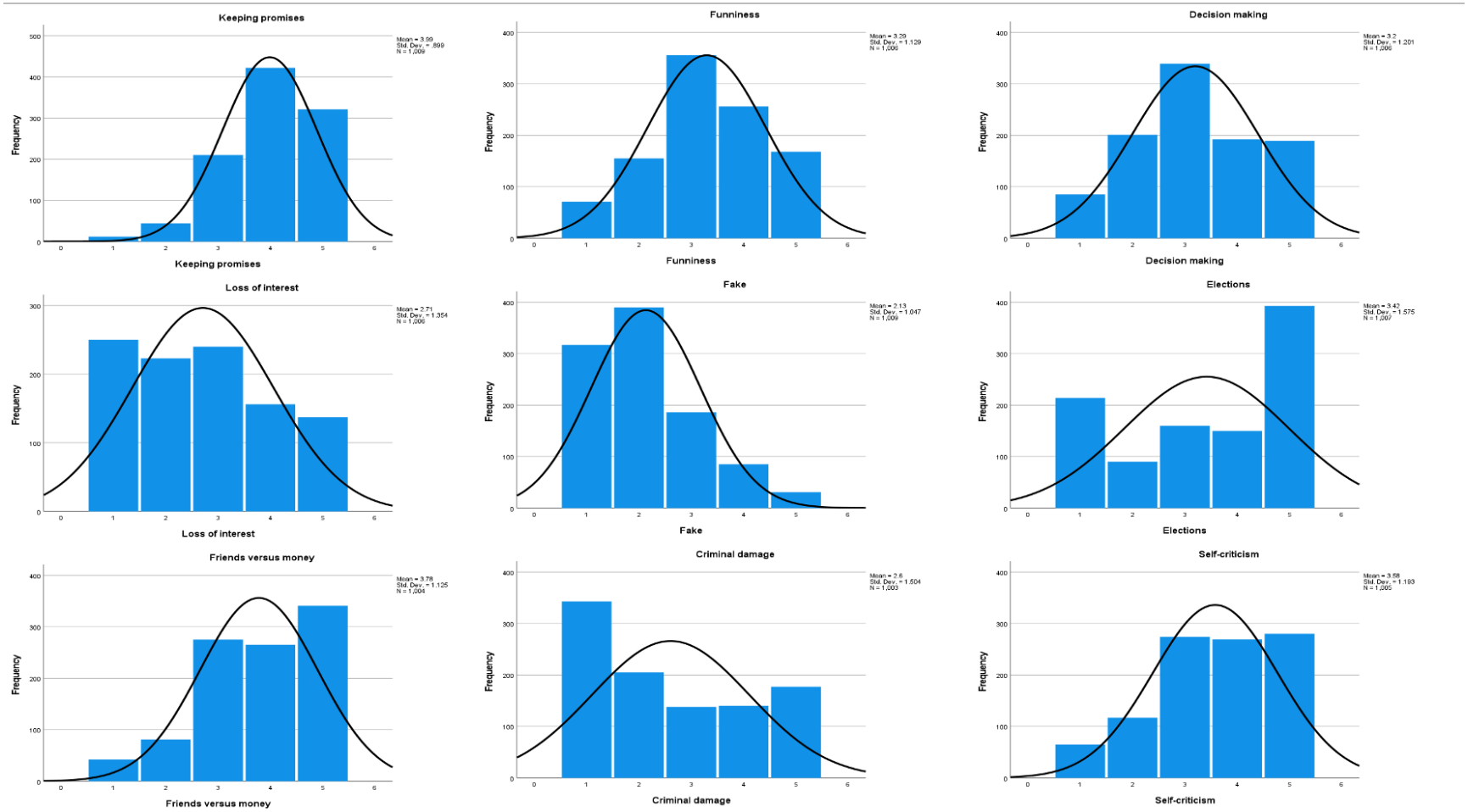

Internet
Frequency
Internet
Biology
Frequency
Biology
Geography
Frequency
Geography
PC
Frequency
PC
Chemistry
Frequency
Chemistry
Foreign languages
Frequency
Foreign languages
Economy Management
Frequency
Economy Management
Reading
Frequency
Reading
Medicine
Frequency
Medicine

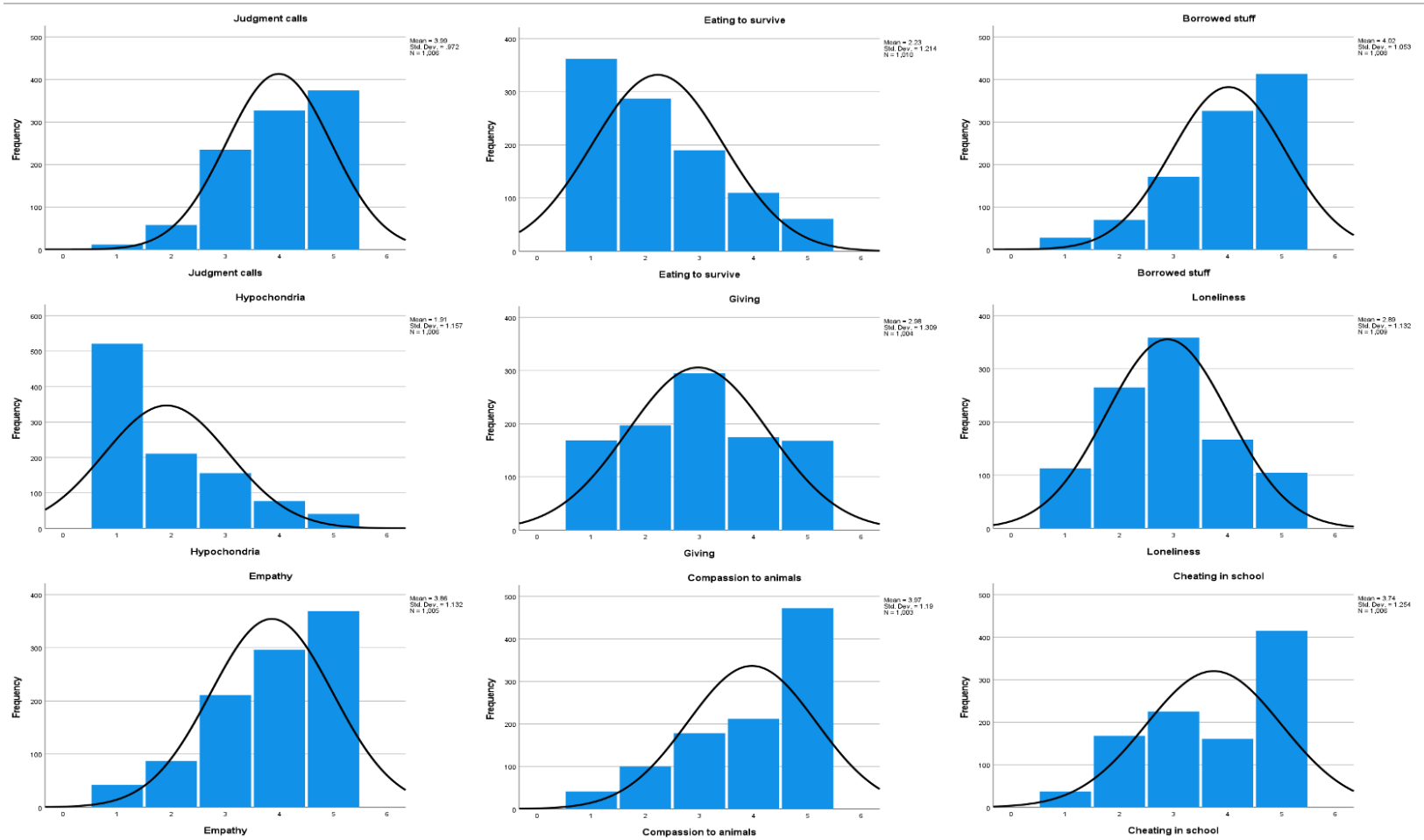

Law
Mean = 2.26
Std. Dev. = 1.243
N = 1,009
Frequency
Law
Religion
Mean = 2.27
Std. Dev. = 1.32
N = 1,007
Frequency
Religion
Musical instruments
Mean = 2.33
Std. Dev. = 1.513
N = 1,009
Frequency
Musical instruments
Cars
Mean = 2.69
Std. Dev. = 1.441
N = 1,006
Frequency
Cars
Countryside, outdoors
Mean = 3.69
Std. Dev. = 1.196
N = 1,003
Frequency
Countryside, outdoors
Writing
Mean = 1.9
Std. Dev. = 1.288
N = 1,004
Frequency
Writing
Art exhibitions
Mean = 2.69
Std. Dev. = 1.322
N = 1,004
Frequency
Art exhibitions
Dancing
Mean = 2.46
Std. Dev. = 1.45
N = 1,007
Frequency
Dancing
Passive sport
Mean = 3.39
Std. Dev. = 1.405
N = 995
Frequency
Passive sport

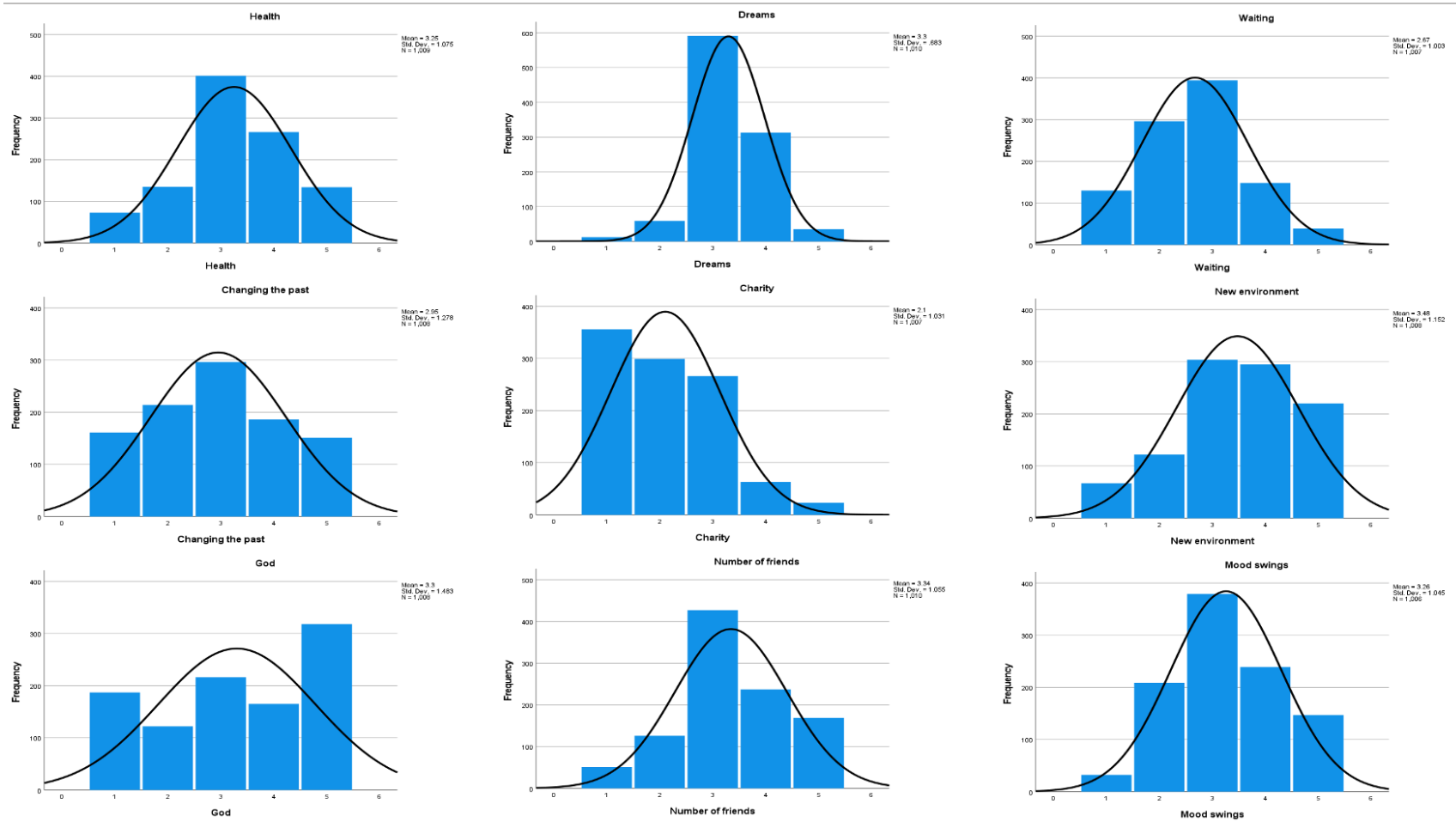

Active sport
Frequency
Active sport
Shopping
Frequency
Shopping
Fun with friends
Frequency
Fun with friends
Gardening
Frequency
Gardening
Science and technology
Frequency
Science and technology
Adrenaline sports
Frequency
Adrenaline sports
Celebrities
Frequency
Celebrities
Theatre
Frequency
Theatre
Pets
Frequency
Pets

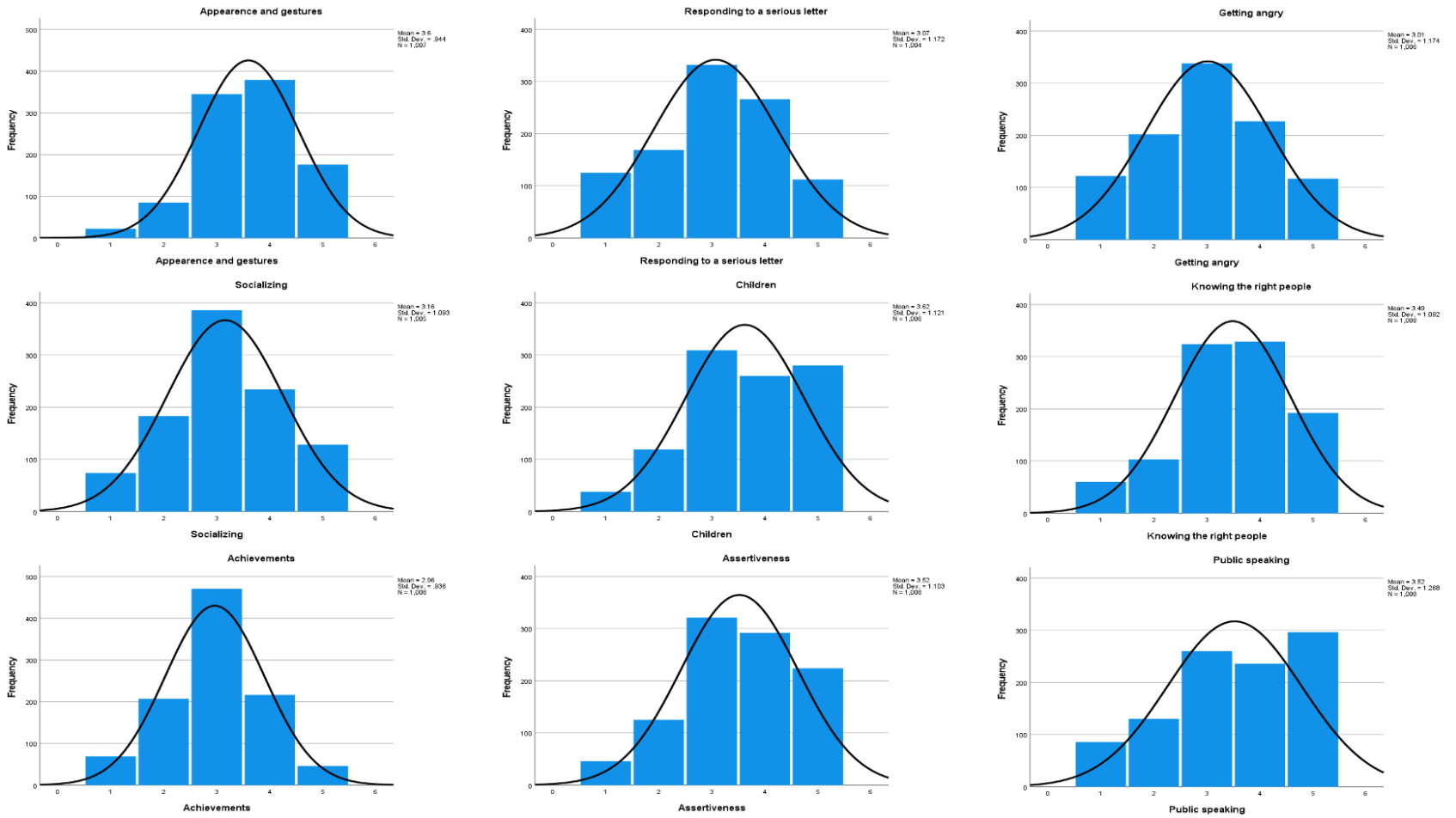

Flying
Mean = 2.06
Std. Dev. = 1.211
N = 1,007
Frequency
Flying
Heights
Mean = 2.62
Std. Dev. = 1.295
N = 1,007
Frequency
Heights
Rats
Mean = 2.41
Std. Dev. = 1.401
N = 1,007
Frequency
Rats
Storm
Mean = 1.97
Std. Dev. = 1.164
N = 1,009
Frequency
Storm
Spiders
Mean = 2.83
Std. Dev. = 1.544
N = 1,005
Frequency
Spiders
Ageing
Mean = 2.58
Std. Dev. = 1.386
N = 1,009
Frequency
Ageing
Darkness
Mean = 2.25
Std. Dev. = 1.255
N = 1,009
Frequency
Darkness
Snakes
Mean = 3.03
Std. Dev. = 1.501
N = 1,010
Frequency
Snakes
Dangerous dogs
Mean = 3.04
Std. Dev. = 1.367
N = 1,009
Frequency
Dangerous dogs

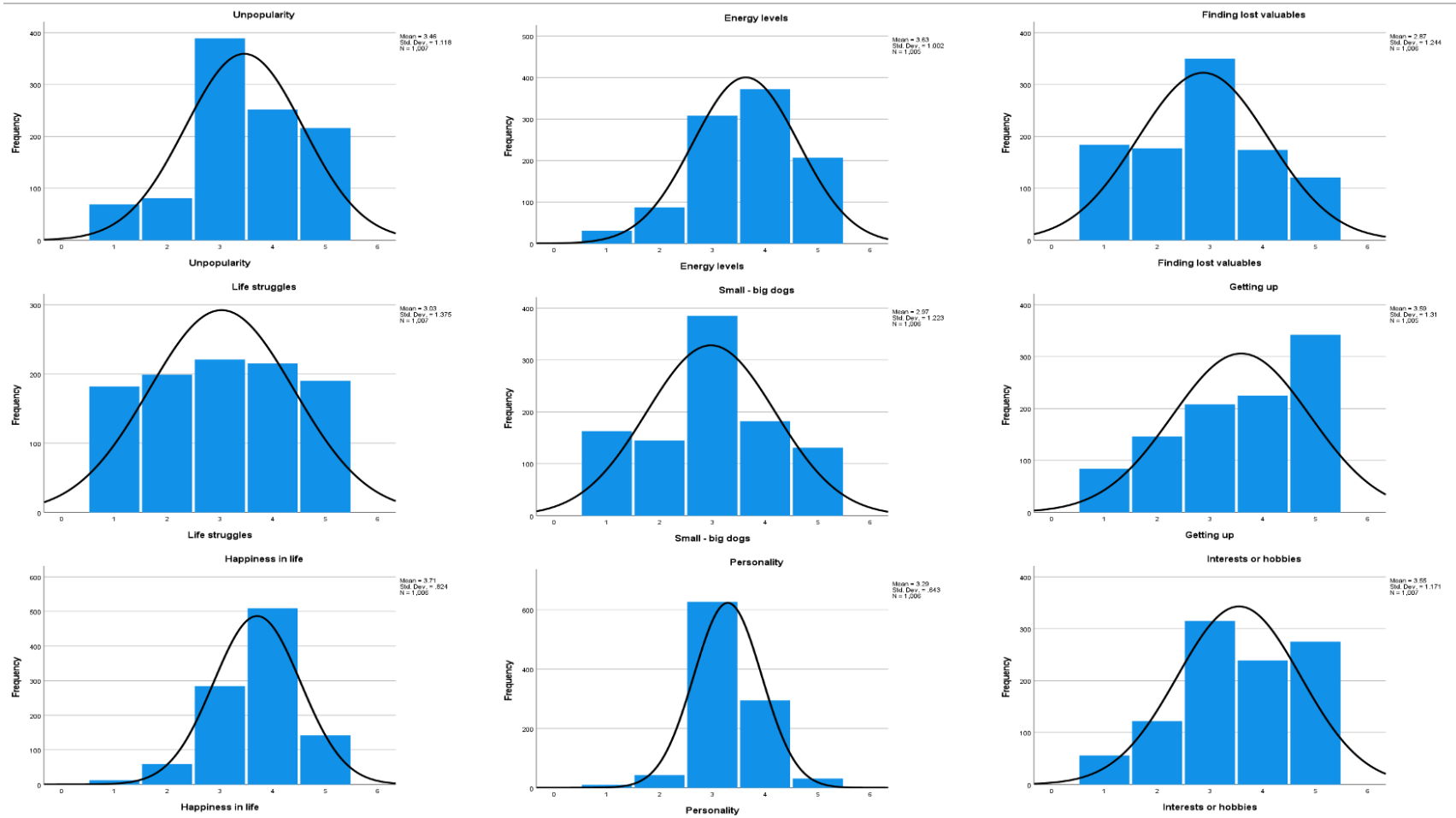

Fear of public speaking
Mean = 2.8
Std. Dev. = 1.215
N = 1,009
Frequency
Fear of public speaking
Prioritising workload
Mean = 2.65
Std. Dev. = 1.221
N = 1,005
Frequency
Prioritising workload
Thinking ahead
Mean = 3.41
Std. Dev. = 1.137
N = 1,007
Frequency
Thinking ahead
Healthy eating
Mean = 3.03
Std. Dev. = .937
N = 1,007
Frequency
Healthy eating
Writing notes
Mean = 3.08
Std. Dev. = 1.408
N = 1,007
Frequency
Writing notes
Final judgement
Mean = 2.65
Std. Dev. = 1.379
N = 1,003
Frequency
Final judgement
Daily events
Mean = 3.07
Std. Dev. = 1.118
N = 1,003
Frequency
Daily events
Workaholism
Mean = 3
Std. Dev. = 1.277
N = 1,005
Frequency
Workaholism
Reliability
Mean = 3.86
Std. Dev. = .834
N = 1,006
Frequency
Reliability

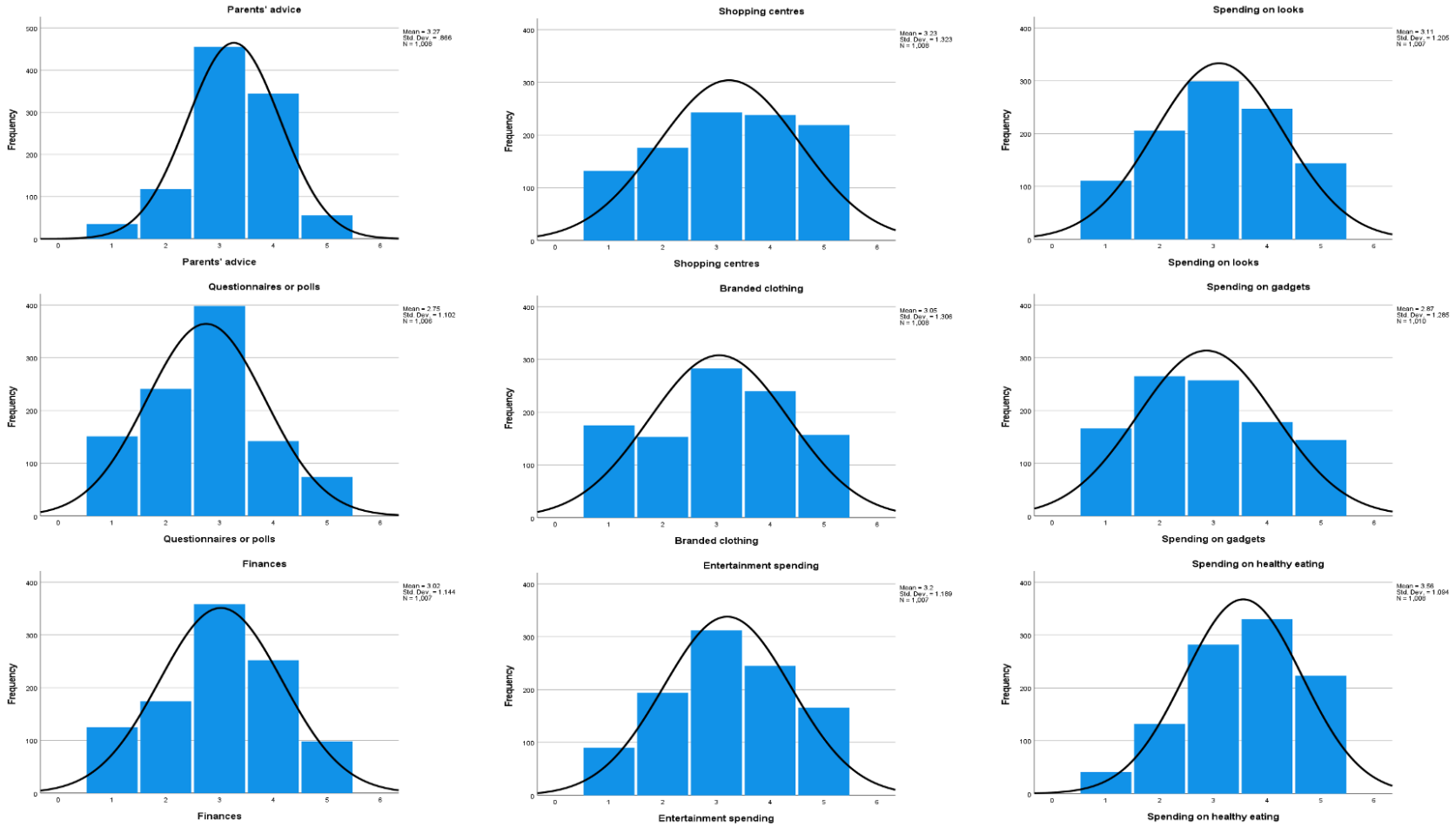

Keeping promises
Frequency
Keeping promises
Funniness
Frequency
Funniness
Decision making
Frequency
Decision making
Loss of interest
Frequency
Loss of interest
Fake
Frequency
Fake
Elections
Frequency
Elections
Friends versus money
Frequency
Friends versus money
Criminal damage
Frequency
Criminal damage
Self-criticism
Frequency
Self-criticism

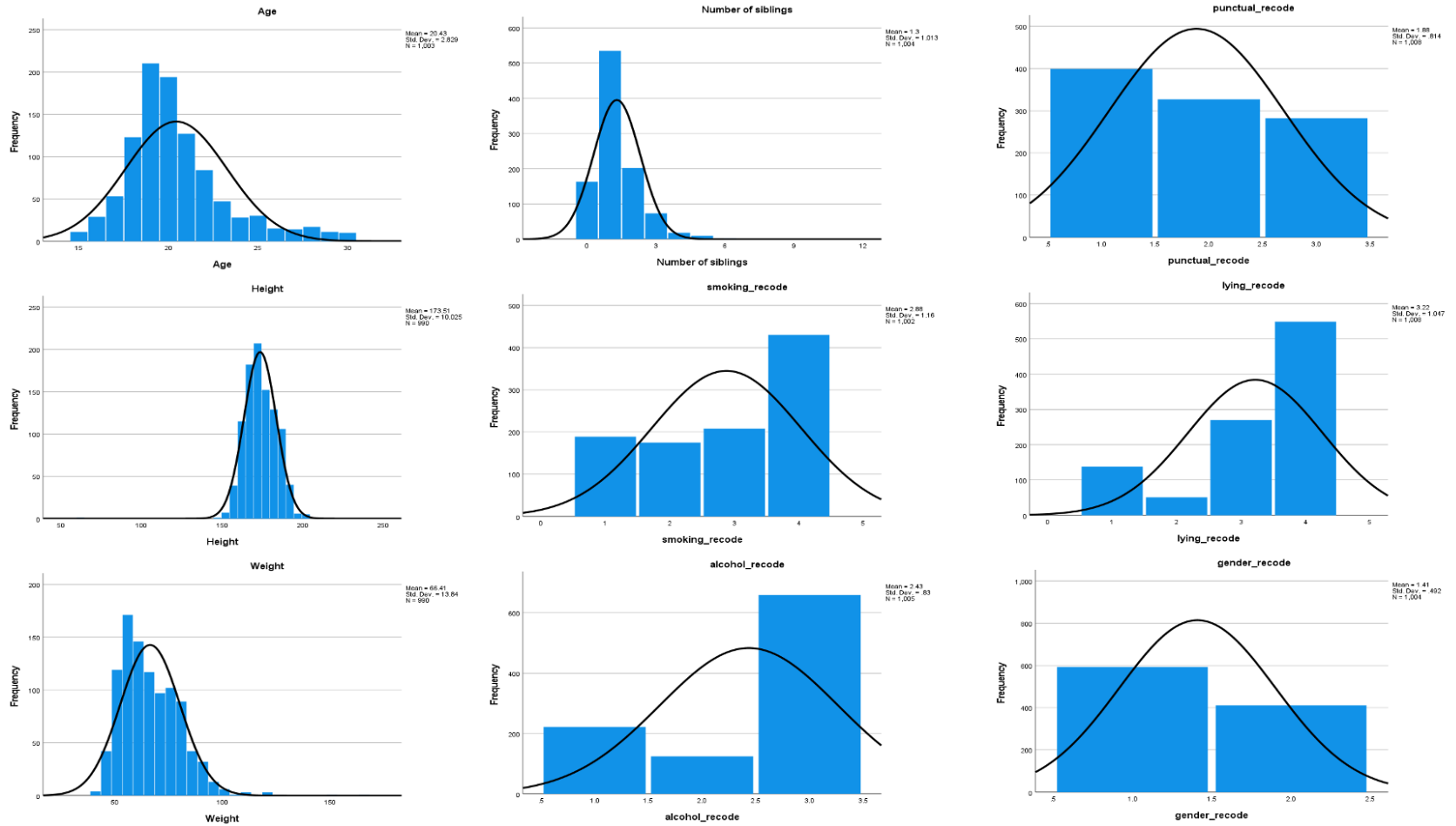

Judgment calls
Frequency
Judgment calls
Eating to survive
Frequency
Eating to survive
Borrowed stuff
Frequency
Borrowed stuff
Hypochondria
Frequency
Hypochondria
Giving
Frequency
Giving
Loneliness
Frequency
Loneliness
Empathy
Frequency
Empathy
Compassion to animals
Frequency
Compassion to animals
Cheating in school
Frequency
Cheating in school

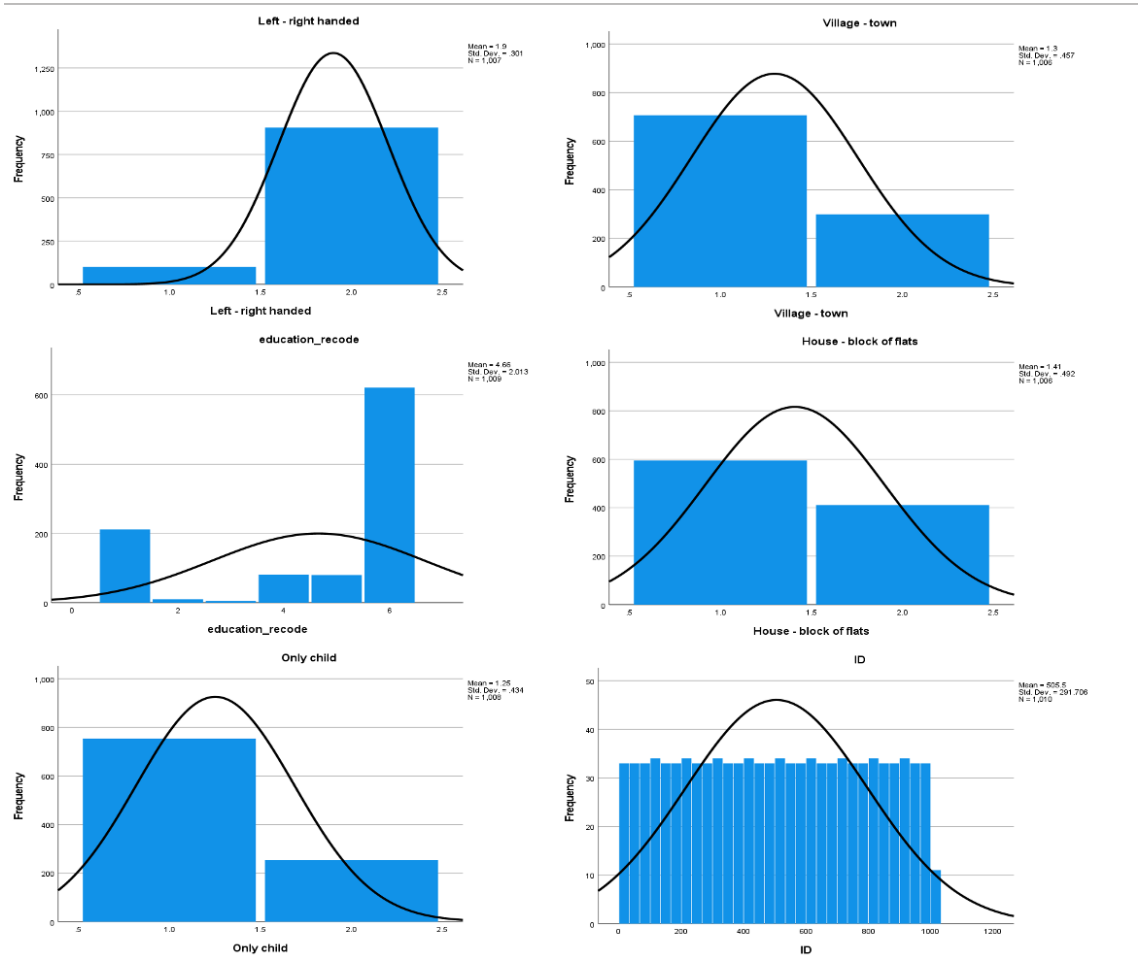

Health
Frequency
Health
Dreams
Frequency
Dreams
Waiting
Frequency
Waiting
Changing the past
Frequency
Changing the past
Charity
Frequency
Charity
New environment
Frequency
New environment
God
Frequency
God
Number of friends
Frequency
Number of friends
Mood swings
Frequency
Mood swings


Appearence and gestures
Frequency
Appearence and gestures
Responding to a serious letter
Frequency
Responding to a serious letter
Getting angry
Frequency
Getting angry
Socializing
Frequency
Socializing
Children
Frequency
Children
Knowing the right people
Frequency
Knowing the right people
Achievements
Frequency
Achievements
Assertiveness
Frequency
Assertiveness
Public speaking
Frequency
Public speaking


Unpopularity
Frequency
Unpopularity
Energy levels
Frequency
Energy levels
Finding lost valuables
Frequency
Finding lost valuables
Life struggles
Frequency
Life struggles
Small - big dogs
Frequency
Small - big dogs
Getting up
Frequency
Getting up
Happiness in life
Frequency
Happiness in life
Personality
Frequency
Personality
Interests or hobbies
Frequency
Interests or hobbies

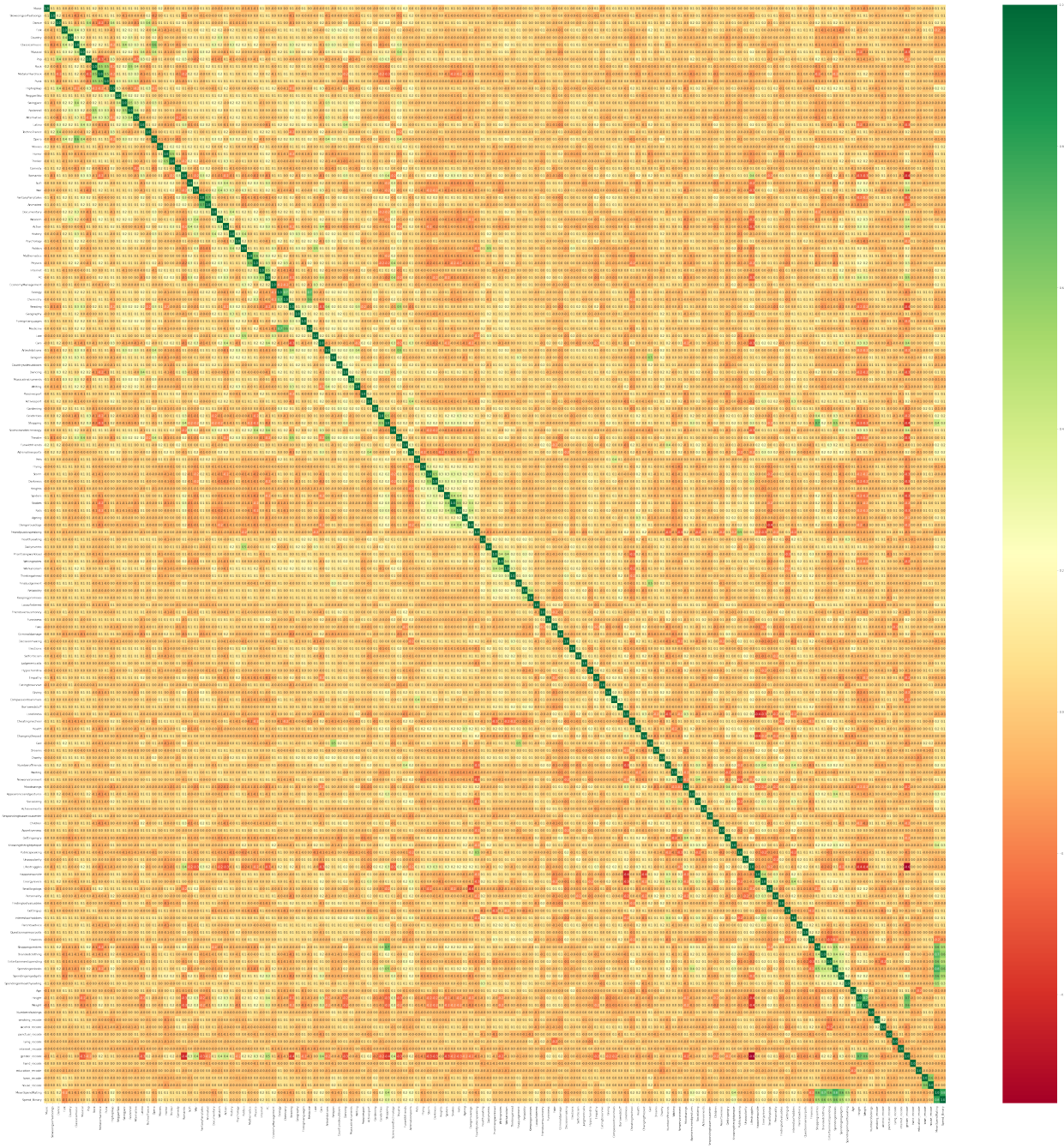

Parents' advice
Frequency
Parents' advice
Shopping centres
Frequency
Shopping centres
Spending on looks
Frequency
Spending on looks
Questionnaires or polls
Frequency
Questionnaires or polls
Branded clothing
Frequency
Branded clothing
Spending on gadgets
Frequency
Spending on gadgets
Finances
Frequency
Finances
Entertainment spending
Frequency
Entertainment spending
Spending on healthy eating
Frequency
Spending on healthy eating


Age
Mean = 20.43
Std. Dev. = 2.829
N = 1,003
Frequency
Age
Number of siblings
Mean = 1.3
Std. Dev. = 1.013
N = 1,004
Frequency
Number of siblings
punctual_recode
Mean = 1.88
Std. Dev. = .814
N = 1,008
Frequency
punctual_recode
Height
Mean = 173.51
Std. Dev. = 10.025
N = 990
Frequency
Height
smoking_recode
Mean = 2.88
Std. Dev. = 1.16
N = 1,002
Frequency
smoking_recode
lying_recode
Mean = 3.22
Std. Dev. = 1.047
N = 1,000
Frequency
lying_recode
Weight
Mean = 66.41
Std. Dev. = 13.84
N = 990
Frequency
Weight
alcohol_recode
Mean = 2.43
Std. Dev. = .83
N = 1,005
Frequency
alcohol_recode
gender_recode
Mean = 1.41
Std. Dev. = .492
N = 1,004
Frequency
gender_recode


Left - right handed
Mean = 1.9
Std. Dev. = .301
N = 1,007
Frequency
Left - right handed
Village - town
Mean = 1.3
Std. Dev. = .457
N = 1,006
Frequency
Village - town
education_recode
Mean = 4.66
Std. Dev. = 2.013
N = 1,009
Frequency
education_recode
House - block of flats
Mean = 1.41
Std. Dev. = .492
N = 1,006
Frequency
House - block of flats
Only child
Mean = 1.26
Std. Dev. = .434
N = 1,009
Frequency
Only child
ID
Mean = 505.5
Std. Dev. = 291.706
N = 1,010
Frequency
ID

Appendix C

Correlation Matrix



Appendix D

Factor Loadings and Naming of Factors

Factor Number	Factor Name	Variable	Description	1	2	3	4	5	6	7
1	Art inclinations	Theatre	Theatre: Not interested 1-2-3-4-5 Very interested (integer)	0.77090151						
1	Art inclinations	Writing	Poetry writing: Not interested 1-2-3-4-5 Very interested (integer)	0.43106093						
1	Art inclinations	Artexhibitions	Art: Not interested 1-2-3-4-5 Very interested (integer)	0.85677107						
1	Art inclinations	Musical	Musicals: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)	0.32014775						
1	Art inclinations	Opera	Opera: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)	0.44193595						
1	Art inclinations	Classicalmusic	Classical: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)	0.29719792						
2	Sciences	Medicine	Medicine: Not interested 1-2-3-4-5 Very interested (integer)		0.8881823					
2	Sciences	Biology	Biology: Not interested 1-2-3-4-5 Very interested (integer)		0.86948544					
2	Sciences	Chemistry	Chemistry: Not interested 1-2-3-4-5 Very interested (integer)		0.78125027					
3	Rock	MetalorHardrock	Metal, Hard rock: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)			0.7930693				
3	Rock	Rock	Rock: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)			0.8181749				
3	Rock	Punk	Punk: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)			0.74944267				
4	Happiness	Energylevels	I am always full of life and energy.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)				-0.3560256			0.3059424
4	Happiness	Loneliness	I feel lonely in life.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)				0.67604366			
4	Happiness	Happinessinlife	I am 100% happy with my life.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)				-0.8113324			
4	Happiness	Changingthepast	I wish I could change the past because of the things I have done.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)				0.42647189			
5	Educational	Documentary	Documentaries: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)					0.75214572		
5	Educational	History	History: Not interested 1-2-3-4-5 Very interested (integer)					0.65352001		
6	Spiritual beliefs	Religion	Religion: Not interested 1-2-3-4-5 Very interested (integer)						0.61654567	
6	Spiritual beliefs	God	I believe in God.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)						0.91296741	
6	Spiritual beliefs	Finaljudgement	I believe that bad people will suffer one day and good people will be rewarded.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)						0.62519379	
7	Physical activity	Interestsorhobbies	I have many different hobbies and interests.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)							0.4677771
7	Physical activity	Adrenalinesports	Adrenaline sports: Not interested 1-2-3-4-5 Very interested (integer)							0.41309101
7	Physical activity	Activesport	Sport at competitive level: Not interested 1-2-3-4-5 Very interested (integer)							0.92496965

Factor Number	Factor Name	Variable	Description	8	9	10	11	12	13	14
8	Pop/ Latino Music	Latino	Latin: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)	0.81451888						
8	Pop/ Latino Music	Pop	Pop: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)	0.30458509						
9	Country musci	Folk	Folk music: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)		0.65829298					
9	Country musci	Country	Country: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)		0.77595933					
10	Social awareness	Law	Law: Not interested 1-2-3-4-5 Very interested (integer)			0.76061659				
10	Social awareness	EconomyManagement	Economy, Management: Not interested 1-2-3-4-5 Very interested (integer)			0.53902295				
11	Computer Technology	Internet	Internet: Not interested 1-2-3-4-5 Very interested (integer)				0.89274559			
11	Computer Technology	PC	PC Software, Hardware: Not interested 1-2-3-4-5 Very interested (integer)				0.65544884			
12	Fantasy Films	FantasyFairytales	Tales: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)					0.91256711		
12	Fantasy Films	Animated	Cartoons: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)					0.78669211		
13	Extrovertness	Newenvironment	I can quickly adapt to a new environment.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)						0.65302872	
13	Extrovertness	Socializing	I enjoy meeting new people.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)						0.86328552	
14	Scary movies	Thriller	Thriller movies: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)							1.0012814
14	Scary movies	Horror	Horror movies: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)							0.5894644

Factor Number	Factor Name	Variable	Description	15	16	17	18	19	20	21
15	Upbeat music	TechnoTrance	Techno, Trance: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)	0.71138569						
15	Upbeat music	Dancing	Dancing: Not interested 1-2-3-4-5 Very interested (integer)							
15	Upbeat music	Dance	Dance, Disco, Funk: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)	0.75346841						
16	Rhythm and Blues	SwingJazz	Swing, Jazz: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)		0.78367373			0.33524362		
16	Rhythm and Blues	Rocknroll	Rock n Roll: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)		0.43407449					
17	Diligence	Writingnotes	I always make a list so I don't forget anything.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)			0.51496841				
17	Diligence	Prioritisingworkload	I try to do tasks as soon as possible and not leave them until last minute.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)			0.6743271				
17	Diligence	Workaholism	I often study or work even in my spare time.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)			0.67256308				
18	Math and Physics	Physics	Physics: Not interested 1-2-3-4-5 Very interested (integer)				0.70125809			
18	Math and Physics	Mathematics	Mathematics: Not interested 1-2-3-4-5 Very interested (integer)				0.75775775			
19	Hip Music	ReggaeSka	Reggae, Ska: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)		0.30216804			0.86992745		
19	Hip Music	HiphopRap	Hip hop, Rap: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)	0.31110858				0.36970852		
20	Observant	Appearenceandgestures	I am well mannered and I look after my appearance.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)						0.52220678	
20	Observant	Dailyevents	I take notice of what goes on around me.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)						0.2630315	
20	Observant	Knowingtherightpeople	I always make sure I connect with the right people.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)						0.60982083	
21	Friendliness	Funwithfriends	Socializing: Not interested 1-2-3-4-5 Very interested (integer)							0.66720666
21	Friendliness	Friendsversusmoney	I would rather have lots of friends than lots of money.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)							0.46021553
21	Friendliness	Numberoffriends	I have lots of friends.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)							0.34771929

Factor Number	Factor Name	Variable	Description	22	23	24	25	26	27	28	29	30
22	Trustworthiness	Keepingpromises	I always keep my promises.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)	0.64516518								
22	Trustworthiness	Reliability	I am reliable at work and always complete all tasks given to me.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)	0.60664159								
23	Animal Person	Compassiontoanimals	I don't like seeing animals suffering.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)		0.80400481							
23	Animal Person	Pets	Pets: Not interested 1-2-3-4-5 Very interested (integer)		0.59630093							
24	Irritability	Moodswings	My moods change quickly.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)			0.54783217						
24	Irritability	Gettingangry	I can get angry very easily.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)			0.76072352						
25	Pretentiousness	Fake	I can be two faced sometimes.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)				0.77599644					
25	Pretentiousness	Funniness	I always try to be the funniest one.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)				0.39674734					
26	Conscientiousness	Selfcriticism	I often think about and regret the decisions I make.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)					0.36559577				
26	Conscientiousness	Thinkingahead	I look at things from all different angles before I go ahead.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)					0.53891936				
26	Conscientiousness	Decisionmaking	I take my time to make decisions.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)					0.61714198				
27	Western and War Films	War	War movies: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)						0.37441088			
27	Western and War Films	Western	Western movies: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)						0.64299063			
28	ScifiAction Movie	Action	Action movies: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)							0.47748717		
28	ScifiAction Movie	Scifi	Sci-fi movies: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)							0.69789634		
29	Righteousness	Unpopularity	I will find a fault in myself if people don't like me.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)								0.46132606	
29	Righteousness	Parentsadvice	I always listen to my parents' advice.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)								0.3492568	
29	Righteousness	Findinglostvaluables	If I find something the doesn't belong to me I will hand it in.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)								0.44371576	
30	Political awareness	Politics	Politics: Not interested 1-2-3-4-5 Very interested (integer)									0.2993
30	Political awareness	Elections	I always try to vote in elections.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)									0.6954

Factor Number	Factor Name	Variable	Description	30	31	32
29	Righteousness	Findinglostvaluables	If I find something the doesn't belong to me I will hand it in.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)			
30	Political awareness	Politics	Politics: Not interested 1-2-3-4-5 Very interested (integer)	0.2993		
30	Political awareness	Elections	I always try to vote in elections.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)	0.6954		
31	Language and Geograp	Geography	Geography: Not interested 1-2-3-4-5 Very interested (integer)		0.30536857	
31	Language and Geograp	Foreignlanguages	Foreign languages: Not interested 1-2-3-4-5 Very interested (integer)		0.42360716	
32	Empathic	Judgmentcalls	I can tell if people listen to me or not when I talk to them.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)			0.563486011
32	Empathic	Empathy	I am emphatetic person.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)			0.538633535
32	Empathic	Charity	I always give to charity.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)			0.249770522

Appendix E

Decision Tree – Metrics Computation in Python

Confusion Matrix:

```
dt_new = confusion_matrix(valid_y, fullClassTree.predict(valid_X_DT))

print("Confusion matrix:\n")
print(dt_new)
tn = dt_new[0][0]
fp = dt_new[0][1]
fn = dt_new[1][0]
tp = dt_new[1][1]

print("\nTotal number of true positives", tp)
print("Total number of false negatives",fn)
print("Total number of false positives",fp)
print("Total number of true negatives",tn)

acc=float(tp+tn)/(tp+tn+fp+fn)

print('\nClassifier Accuracy: %.2f%%' % (acc * 100))

tpr = float(tp)/(tp+fn)

print('True Positive Rate (TPR/Recall/Sensitivity): %.2f%%' % (tpr *
100))

specificity = float (tn)/(tn+fp)

print ("True Negative Rate (TNR/Specificity/selectivity):%.2f%%" %
(specificity*100))

fpr = float(fp)/(fp+tn)
print("False Positive Rate (FPR): %.2f%%" % (fpr * 100))

fnr = fn/ (fn+ tp)
print("False Negative Rate (FNR): %.2f%%" % (fnr*100))

precision=float(tp)/(tp+fp)
print("Precision/Positive Predictive value:%.2f%%" %(precision*100))

fScore = 2*((precision*tpr)/(precision+tpr))
print("F1-Score: %.2f%%" %(fScore*100))
```

Mean Squared Error

```
# evaluate model using MSE
display('FullDTModelMSE:',
mean_squared_error(valid_y,fullClassTree.predict(valid_X_DT)))
```

ROC-AUC

```
y_pred_probafulldt = fullClassTree.predict_proba(valid_X_DT)[::,1]
#calculate probabilities high spender (label=1)

fpr, tpr, _ = roc_curve(valid_y, y_pred_probafulldt) # calculate tpr
and fpr values

auc = roc_auc_score(valid_y, y_pred_probafulldt) #calculate auc value


plt.plot(fpr,tpr,label="test_data(AUC= %0.2f)" % auc, linewidth = 4)
#plot ROC curve
plt.legend(prop={'size':12},loc='best') #set the legend properties
plt.title('\nROC Plot - Full DT',fontsize = 16) #title
plt.xlabel('False positive rate', fontsize = 16) #x and y labels
plt.ylabel('True positive rate', fontsize = 16)
plt.show() #display the plot
```

Appendix F

Grid Search – Metrics Computation in Python

Confusion Matrix:

```
bt_new = confusion_matrix(valid_y, besttree.predict(valid_X_DT))

print("Confusion matrix:\n")
print(bt_new)
tn = bt_new[0][0]
fp = bt_new[0][1]
fn = bt_new[1][0]
tp = bt_new[1][1]

print("\nTotal number of true positives", tp)
print("Total number of false negatives",fn)
print("Total number of false positives",fp)
print("Total number of true negatives",tn)

acc=float(tp+tn)/(tp+tn+fp+fn)
print('\nClassifier Accuracy: %.2f%%' % (acc * 100))

tpr = float(tp)/(tp+fn)
print('True Positive Rate (TPR/Recall/Sensitivity): %.2f%%' % (tpr *
100))

specificity = float (tn)/(tn+fp)
print ("True Negative Rate (TNR/Specificity/selectivity):%.2f%%" %
(specificity*100))

fpr = float(fp)/(fp+tn)
print("False Positive Rate (FPR): %.2f%%" % (fpr * 100))

fnr = fn/ (fn+ tp)
print("False Negative Rate (FNR): %.2f%%" % (fnr*100))

precision=float(tp)/(tp+fp)
print("Precision/Positive Predictive value:%.2f%%" %(precision*100))

fScore = 2*((precision*tpr)/(precision+tpr))
print("F1-Score: %.2f%%" %(fScore*100))
```

Mean Squared Error

```
# evaluate model using MSE
display('BestDTModelMSE:',
mean_squared_error(valid_y,besttree.predict(valid_X_DT)))
```

ROC-AUC

```
y_pred_probabt = besttree.predict_proba(valid_X_DT)[::,1] #calculate
probabilities for high spender YES (label=1)

fpr, tpr, _ = roc_curve(valid_y, y_pred_probabt) # calculate tpr and
fpr values

auc = roc_auc_score(valid_y, y_pred_probabt) #calculate auc value


plt.plot(fpr,tpr,label="test_data(AUC= %0.2f)" % auc, linewidth = 4)
#plot ROC curve
plt.legend(prop={'size':12},loc='best') #set the legend properties
plt.title('\nROC Plot - Best DT',fontsize = 16) #title
plt.xlabel('False positive rate', fontsize = 16) #x and y labels
plt.ylabel('True positive rate', fontsize = 16)
plt.show() #display the plot
```

Appendix G

Full Logistic Regression – Metrics Computation in Python

Confusion Matrix:

```
full_new = confusion_matrix(valid_y, logitreg_pred)

print("Confusion matrix:\n")
print(full_new)
tn = full_new[0][0]
fp = full_new[0][1]
fn = full_new[1][0]
tp = full_new[1][1]

print("\nTotal number of true positives", tp)
print("Total number of false negatives",fn)
print("Total number of false positives",fp)
print("Total number of true negatives",tn)

acc=float(tp+tn)/(tp+tn+fp+fn)
print('\nClassifier Accuracy: %.2f%%' % (acc * 100))

tpr = float(tp)/(tp+fn)
print('True Positive Rate (TPR/Recall/Sensitivity): %.2f%%' % (tpr *
100))

specificity = float (tn)/(tn+fp)
print ("True Negative Rate (TNR/Specificity/selectivity):%.2f%%" %
(specificity*100))

fpr = float(fp)/(fp+tn)
print("False Positive Rate (FPR): %.2f%%" % (fpr * 100))

fnr = fn/ (fn+ tp)
print("False Negative Rate (FNR): %.2f%%" % (fnr*100))

precision=float(tp)/(tp+fp)
print("Precision/Positive Predictive value:%.2f%%" %(precision*100))

fScore = 2*((precision*tpr)/(precision+tpr))
print("F1-Score: %.2f%%" %(fScore*100))
```

Mean Squared Error:

```
# evaluate model using MSE
display('FullModelMSE:', mean_squared_error(valid_y,logitreg_pred))
```

ROC-AUC:

```
y_pred_probafull = logitreg_prob[::,1] #calculate probabilities for
high spender YES (label=1)

fpr, tpr, _ = roc_curve(valid_y, y_pred_probafull) # calculate tpr and
fpr values

auc = roc_auc_score(valid_y, y_pred_probafull) #calculate auc value


plt.plot(fpr,tpr,label="valid_data(AUC= %0.2f)" % auc, linewidth = 4)
#plot ROC curve
plt.legend(prop={'size':12},loc='best') #set the legend properties
plt.title('\nROC Plot - Full',fontsize = 16) #title
plt.xlabel('False positive rate', fontsize = 16) #x and y labels
plt.ylabel('True positive rate', fontsize = 16)
plt.show() #display the plot
```

Appendix H

Forward Step Feature Selection – Metrics Computation in Python

Confusion Matrix:

```
fwdnew = confusion_matrix(valid_y, y_valid_predfwd)

print("Confusion matrix:\n")
print(fwdnew)
tn = fwdnew[0][0]
fp = fwdnew[0][1]
fn = fwdnew[1][0]
tp = fwdnew[1][1]

print("\nTotal number of true positives",tp)
print("Total number of false negatives",fn)
print("Total number of false positives",fp)
print("Total number of true negatives",tn)

acc=float(tp+tn)/(tp+tn+fp+fn)
print('\nClassifier Accuracy: %.2f%%' % (acc * 100))

tpr = float(tp)/(tp+fn)
print('True Positive Rate (TPR/Recall/Sensitivity): %.2f%%' % (tpr *
100))

specificity = float (tn)/(tn+fp)
print("True Negative Rate (TNR/Specificity/selectivity):%.2f%%" %
(specificity*100))

fpr = float(fp)/(fp+tn)
print("False Positive Rate (FPR): %.2f%%" % (fpr * 100))

fnr = fn/ (fn+ tp)
print("False Negative Rate (FNR): %.2f%%" % (fnr*100))

precision=float(tp)/(tp+fp)
print("Precision/Positive Predictive value:%.2f%%" %(precision*100))

fScore = 2*((precision*tpr)/(precision+tpr))
print("F1-Score: %.2f%%" %(fScore*100))
```

Mean Squared Error:

```
# evaluate model using MSE
display('FwdModelMSE:', mean_squared_error(valid_y,y_valid_predfwd))
```

ROC-AUC:

```
y_pred_probafwd = logitreg_probfwd[::,1] #calculate probabilities for
high spender YES (label=1)

fpr, tpr, _ = roc_curve(valid_y, y_pred_probafwd) # calculate tpr and
fpr values

auc = roc_auc_score(valid_y, y_pred_probafwd) #calculate auc value


plt.plot(fpr,tpr,label="valid_data(AUC= %0.2f)" % auc, linewidth = 4)
#plot ROC curve
plt.legend(prop={'size':12},loc='best') #set the legend properties
plt.title('\nROC plot - Forward',fontsize = 16) #title
plt.xlabel('False positive rate', fontsize = 16) #x and y labels
plt.ylabel('True positive rate', fontsize = 16)
plt.show() #display the plot
```

Appendix I

Backward Elimination Feature Selection – Metrics Computation in Python

Confusion Matrix:

```
bwd_new = confusion_matrix(valid_y, y_valid_predbwd)

print("Confusion matrix:\n")
print(bwd_new)
tn = bwd_new[0][0]
fp = bwd_new[0][1]
fn = bwd_new[1][0]
tp = bwd_new[1][1]

print("\nTotal number of true positives", tp)
print("Total number of false negatives",fn)
print("Total number of false positives",fp)
print("Total number of true negatives",tn)

acc=float(tp+tn)/(tp+tn+fp+fn)
print('\nClassifier Accuracy: %.2f%%' % (acc * 100))

tpr = float(tp)/(tp+fn)
print('True Positive Rate (TPR/Recall/Sensitivity): %.2f%%' % (tpr *
100))

specificity = float (tn)/(tn+fp)
print("True Negative Rate (TNR/Specificity/selectivity):%.2f%%" %
(specificity*100))

fpr = float(fp)/(fp+tn)
print("False Positive Rate (FPR): %.2f%%" % (fpr * 100))

fnr = fn/ (fn+ tp)
print("False Negative Rate (FNR): %.2f%%" % (fnr*100))

precision=float(tp)/(tp+fp)
print("Precision/Positive Predictive value:%.2f%%" %(precision*100))

fScore = 2*((precision*tpr)/(precision+tpr))
print("F1-Score: %.2f%%" %(fScore*100))
```

Mean Squared Error:

```
# evaluate model using MSE
display('BwdModelMSE:', mean_squared_error(valid_y,y_valid_predbwd))
```

ROC-AUC:

```
y_pred_probabwd = logitreg_probbwd[::,1] #calculate probabilities for
high spender YES (label=1)

fpr, tpr, _ = roc_curve(valid_y, y_pred_probabwd) # calculate tpr and
fpr values

auc = roc_auc_score(valid_y, y_pred_probabwd) #calculate auc value

plt.plot(fpr,tpr,label="valid_data(AUC= %0.2f)" % auc, linewidth = 4)
#plot ROC curve
plt.legend(prop={'size':12},loc='best') #set the legend properties
plt.title('\nROC Plot - Backward ',fontsize = 16) #title
plt.xlabel('False positive rate', fontsize = 16) #x and y labels
plt.ylabel('True positive rate', fontsize = 16)
plt.show() #display the plot
```

Appendix J

Stepwise Feature Selection – Metrics Computation in Python

Confusion Matrix:

```
stpwise_new = confusion_matrix(valid_y, y_valid_predsw)

print("Confusion matrix:\n")
print(stpwise_new)
tn = stpwise_new[0][0]
fp = stpwise_new[0][1]
fn = stpwise_new[1][0]
tp = stpwise_new[1][1]

print("\nTotal number of true positives", tp)
print("Total number of false negatives",fn)
print("Total number of false positives",fp)
print("Total number of true negatives",tn)

acc=float(tp+tn)/(tp+tn+fp+fn)

print('\nClassifier Accuracy: %.2f%%' % (acc * 100))
tpr = float(tp)/(tp+fn)

print('True Positive Rate (TPR/Recall/Sensitivity): %.2f%%' % (tpr *
100))

specificity = float (tn)/(tn+fp)
print("True Negative Rate (TNR/Specificity/selectivity):%.2f%%" %
(specificity*100))

fpr = float(fp)/(fp+tn)
print("False Positive Rate (FPR): %.2f%%" % (fpr * 100))

fnr = fn/ (fn+ tp)
print("False Negative Rate (FNR): %.2f%%" % (fnr*100))

precision=float(tp)/(tp+fp)
print("Precision/Positive Predictive value:%.2f%%" %(precision*100))

fScore = 2*((precision*tpr)/(precision+tpr))
print("F1-Score: %.2f%%" %(fScore*100))
```

Mean Squared Error:

```
# evaluate model using MSE
display('SwModelMSE:', mean_squared_error(valid_y,y_valid_predsw))
```

ROC-AUC:

```
y_pred_probasw = logitreg_probsw[::,1] #calculate probabilities for
high spender YES (label=1)

fpr, tpr, _ = roc_curve(valid_y, y_pred_probasw) # calculate tpr and
fpr values

auc = roc_auc_score(valid_y, y_pred_probasw) #calculate auc value


plt.plot(fpr,tpr,label="valid_data(AUC= %0.2f)" % auc, linewidth = 4)
#plot ROC curve
plt.legend(prop={'size':12},loc='best') #set the legend properties
plt.title('\nROC Plot - Stepwise',fontsize = 16) #title
plt.xlabel('False positive rate', fontsize = 16) #x and y labels
plt.ylabel('True positive rate', fontsize = 16)
plt.show() #display the plot
```

Appendix K

Cluster Analysis

```
/*With gender and education, AIC

TWOSTEP CLUSTER
  /CATEGORICAL VARIABLES=Gender Education
  /CONTINUOUS VARIABLES=Art Happiness Spritualbeliefs Physicalactivity
Socialawareness
    Computertechnology Extrovert Diligence Trendy Friendliness
Trustworthiness Loveforanimals
    Irritability Pretentiousness Conscientiousness Righteousness
PoliticalAwareness Empathy2
  /DISTANCE LIKELIHOOD
  /NUMCLUSTERS AUTO 15 AIC
  /HANDLENOISE 0
  /MEMALLOCATE 64
  /CRITERIA INITHRESHOLD(0) MXBRANCH(8) MXLEVEL(3)
  /VIEWMODEL DISPLAY=YES.

/*With gender and education, n_cluster = 5

TWOSTEP CLUSTER
  /CATEGORICAL VARIABLES=Gender Education
  /CONTINUOUS VARIABLES=Art Happiness Spritualbeliefs Physicalactivity
Socialawareness
    Computertechnology Extrovert Diligence Trendy Friendliness
Trustworthiness Loveforanimals
    Irritability Pretentiousness Conscientiousness Righteousness
PoliticalAwareness Empathy2
  /DISTANCE LIKELIHOOD
  /NUMCLUSTERS FIXED=5
  /HANDLENOISE 0
  /MEMALLOCATE 64
  /CRITERIA INITHRESHOLD(0) MXBRANCH(8) MXLEVEL(3)
  /VIEWMODEL DISPLAY=YES.


/*With gender only, AIC

TWOSTEP CLUSTER
  /CATEGORICAL VARIABLES=Gender
  /CONTINUOUS VARIABLES=Art Happiness Spritualbeliefs Physicalactivity
Socialawareness
    Computertechnology Extrovert Diligence Trendy Friendliness
Trustworthiness Loveforanimals
    Irritability Pretentiousness Conscientiousness Righteousness
PoliticalAwareness Empathy2
  /DISTANCE LIKELIHOOD
  /NUMCLUSTERS AUTO 15 AIC
  /HANDLENOISE 0
  /MEMALLOCATE 64
  /CRITERIA INITHRESHOLD(0) MXBRANCH(8) MXLEVEL(3)
  /VIEWMODEL DISPLAY=YES.
```

```
/*With gender only, n_cluster = 5

TWOSTEP CLUSTER
  /CATEGORICAL VARIABLES=Gender
  /CONTINUOUS VARIABLES=Art Happiness Spritualbeliefs Physicalactivity
Socialawareness
    Computertechnology Extrovert Diligence Trendy Friendliness
Trustworthiness Loveforanimals
    Irritability Pretentiousness Conscientiousness Righteousness
PoliticalAwareness Empathy2
  /DISTANCE LIKELIHOOD
  /NUMCLUSTERS FIXED=5
  /HANDLENOISE 0
  /MEMALLOCATE 64
  /CRITERIA INITHRESHOLD(0) MXBRANCH(8) MXLEVEL(3)
  /VIEWMODEL DISPLAY=YES
  /SAVE VARIABLE=Segment.
```

REFERENCES

United Nations. (17 July, 2022). *Youth*. Retrieved July, 2022, from United Nations: https://www.un.org/en/global-issues/youth

Minns, A. (18 July, 2022). *Rural Neets in Slovakia*. Retrieved from rnyobservatory.eu: https://rnyobservatory.eu/web/National-Reports/NR-SLOVAKIA-09-19.pdf

ecommerceDB. (18 July, 2022). *The eCommerce Market in Slovakia*. Retrieved from ecommerceDB: https://ecommercedb.com/en/markets/sk/all

PwC. (25 October, 2018). *Slovak women earn 18% less than men – how to achieve equal pay?* Retrieved from PwC: https://www.pwc.com/sk/en/current-press-releases/slovak-women-earn.html

eCommerce360. (11 August, 2022). *Howard Sheth Model of Consumer Behavior*. Retrieved from eCommerce360: https://www.marketing360.in/howard-sheth-model-of-consumer-behavior/