

TO ANALYSE THE EFFECT OF COVID-19 ON INDIVIDUALS' PERSONALITY TRAITS AND THEIR SPENDING PATTERNS

FINAL REPORT

MENTORED BY:

ANIMESH TIWARI

SUBMITTED BY:

ANIRUDH REDDY ITIKELA

JAGTHAP ROHAN KUMAR

KAKARLA SRI RANGANADH

MEGHANA ANANTANENI

SHREYASHI TIWARI

ACKNOWLEDGEMENT

It gives us immense pleasure in bringing out this synopsis of the project entitled "To analyse the effects of COVID-19 on individuals' personality traits and their spending behaviour"

Firstly we would like to thank our centre head Mr.Pushkar Shah and Great learning, Hyderabad for providing us the opportunity to work on this project.

We would like to extend our sincere gratitude to our mentor Mr.Animesh Tiwari whose guidance, encouragement and suggestions have contributed immensely to our project.

We are grateful to our teacher assistant Mrs.Kanchan Wadhava for extending our survey to various platforms.

We would also like to express our gratitude to our friends and respondents for the support and willingness to spend some time with us to fill in the questionnaires. Last but not the least, we are thankful to everyone who helped us directly or indirectly in making this project a successful one.

Contents

Capstone Project–Final Report	4
1. Background Research and Abstract	4
2. Introduction to Psychometrics	6
3. Literature Survey	8
4. Data Introduction	9
4.1. Data Dictionary	9
4.2. Pre-Processing	11
5. Data Exploration (EDA)	20
5.1. Relationship between variables	20
5.2. Multi-Collinearity	29
5.3. Distribution of variables	30
5.4. Presence of outliers and its treatment	30
5.5. Statistical significance of variables	30
5.6. Class Imbalance and its treatment	32
6. Feature Engineering	34
6.1. Scaling	34
6.2. Feature Selection	34
6.3. Dimensionality reduction	35
7. Supervised Learning(Classification)	36
7.1. Base Model	36
7.2. Other Models	37
7.3 Final Model(Gradient Boost)	37
8. Unsupervised Learning (Clustering)	39
8.1. K-Means	39
8.2 Hierarchical(Agglomerative) Clustering	40
8.3 K-Medoids	41
9. Web Application	43
10. Results	44
11. Conclusion	44
12. Drawbacks	45
13. Vision	45
14. Applications	45
15. Executive Summary	46
16. References	48

Capstone Project–Final Report

1. Background Research and Abstract

The global pandemic that we are facing today is an unusual and surreal experience and it has moved everyone to levels unthinkable concerning various aspects of our lives. We felt that this moment of change would be interesting to capture and analyse and started thinking in this direction.

The pandemic affected our physical and mental health alike, irrespective of the geographical boundaries. While physical health is being analysed, quantified and treated, we felt it is the change regarding the human psyche that will be interesting enough to capture and analyse. Since the pandemic changed the way we work, socialize and conduct our day to day activities, we felt, it will also be interesting to analyse and observe the patterns in which people are spending and their priorities regarding the same.

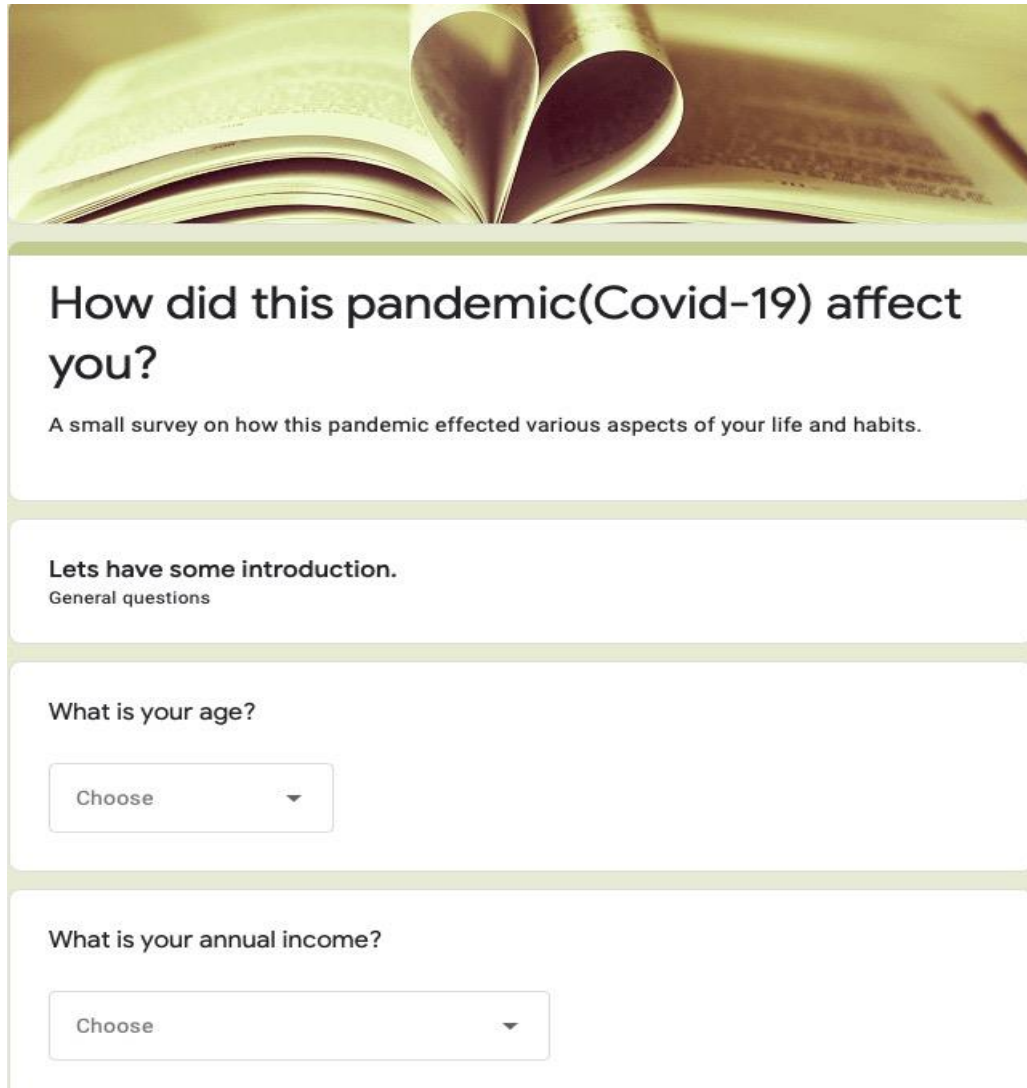
Therefore, the main idea of our project is to analyse the effect of COVID-19 on individual's personality traits and their spending patterns. In this regard, we want to use the statistical and data science tools that we have learned to achieve the following objectives:

- Do exploratory data analysis and observe interesting changes in personality traits and spending behaviour.
- Looking at this as a prediction problem (especially a classification problem) and predicting what spending category (high, medium, low) a person with a set of characteristics will belong to.
- Considering this as a clustering problem and observe, what were the dominant personalities concerning interests, preferences and spending patterns and how did it change now.
- Create an insightful dashboard that will be socially and commercially informative.

So, to achieve the purpose of our problem statement and the respective objectives, we need data that would facilitate this process. To collect data, we have chosen the medium of a survey, where we asked a few questions that would support the kind of analysis we want to do. The following are the list of questions that are segmented into four categories, personality traits, interests, health habits, and spending habits:
survey link: <https://forms.gle/iYq7fv5hS8xaSZGb6>.

Data description link: <https://drive.google.com/file/d/1MebIsDp1F5Ryq1-oVkqPSfwx-RVPCeLK/view?usp=sharing>

The survey recorded 551 responses and contained a total of 46 questions. Finally, from the final responses, two datasets containing the information of Before COVID and Present and of sizes 551×43 and 551×45 were obtained. The dataset consists of all the categorical features which are mostly in Likert scale.



How did this pandemic(Covid-19) affect you?

A small survey on how this pandemic effected various aspects of your life and habits.

Lets have some introduction.
General questions

What is your age?

Choose ▼

What is your annual income?

Choose ▼

Fig(1.1)

Fig(1.1) shows us the questionnaires asked in the survey.

2.Introduction to Psychometrics

Psychometrics is that area of psychology that specializes in how to measure what we talk and think about. It is how to assign numbers to observations in a way that best allows us to summarize our observations in order to advance our knowledge. Although in particular it is the study of how to measure psychological constructs, the techniques of psychometrics are applicable to most problems in measurement. The measurement of intelligence, extraversion, severity of crimes, or even batting averages in baseball are all grist for the psychometric mill. Any set of observations that are not perfect exemplars of the construct of interest is open to questions of reliability and validity and to psychometric analysis.

Psychometricians have developed a number of different measurement theories. These include classical test theory (CTT) and item response theory (IRT). An approach which seems mathematically to be similar to IRT but also quite distinctive, in terms of its origins and features, is represented by the Rasch model for measurement. The development of the Rasch model, and the broader class of models to which it belongs, was explicitly founded on requirements of measurement in the physical sciences. Thus, to relate psychometrics with analytics, Computational Psychometrics was adopted.

Computational Psychometrics is an interdisciplinary field fusing theory-based psychometrics, learning and cognitive sciences, and data-driven AI-based computational models as applied to large-scale/high-dimensional learning, assessment, biometric, or psychological data. Computational psychometrics is frequently concerned with providing actionable and meaningful feedback to individuals based on measurement and analysis of individual differences as they pertain to specific areas of enquiry.[01]

The relatively recent availability of large-scale psychometric data in accessible formats, alongside the rapid increase in CPU processing power, widespread accessibility and application of cluster and cloud computing, and the development of increasingly sensitive instruments for collecting biometric information has allowed big-data analytical and computational methods to expand the scale and scope of traditional psychometric areas of enquiry and modelling.

Computational psychometrics incorporates both theoretical and applied components ranging from item response theory, classical test theory, and Bayesian approaches to modelling knowledge acquisition and discovery of network psychometric models. Computational psychometrics studies the computational basis of learning and measurement of traits, such as skills, knowledge, abilities, attitudes, and personality traits via mathematical modelling, intelligent learning and assessment virtual systems, and computer simulation of large-scale, complex data which traditional psychometric approaches are ill-equipped to handle. Recent

investigations into these hard to measure constructs include work on collaborative problem solving, teamwork, and decision making, among others.

Computational psychometrics is also related to the study of social complexity. Concepts such as complex systems and emergence have been considered in the study of team assembly and performance. In psychological and medical research it is focused on computational models based on technology enhanced-experimental results. Active areas of enquiry include cognitive, emotional, behavioural, diagnostic, and mental health issues. A computational psychometrics approach in this capacity frequently makes use of emerging capabilities such as biometric and multimodal sensors, virtual and augmented reality, as well as affective and wearable computing technologies. Another application of Computational psychometrics comes from marketing where population is segmented, and insights are derived for their additive and interactive effects on various products.

3.Literature Survey

The research papers related to the study of personality traits using different methodologies produced in the previous years are described below:

Charu Nath et al. (2011) have worked on behavioural analysis using various Clustering methods. The objective of the work was to cluster the final year students based on their Behaviour and Personality Traits. The dataset used by the authors was collected with the help of conducting interviews of the Students. The different Clustering methods like K-means, Hierarchical clustering were used to cluster the Students. Some mixed or overlapping clusters were obtained due to the data collection techniques but the authors concluded that the clustering techniques can favourably be applied to the field of behavioural analysis and therefore can prove to be of great use to human-intensive institutes like IT industries and educational institutes.[02]

Alexandros Ladas et al. (2013) have worked to extract Behavioural Groups by using simple clustering techniques that can potentially reveal aspects of the Personalities for their members. The dataset had 58 attributes and it contains information about 70000 clients who contacted the service between the years 2004 and 2008 in order to require advice about how they can overcome their debts. The dataset has 18% of missing value but since the dataset was huge and missing values were present in specific clients, they choose to drop those information. K-means clustering and Clara was used to cluster the dataset. The authors concluded that it is possible to extract information regarding the Personality of individuals from similar datasets by using even simplistic data mining techniques. Also they clustered their dataset as Selfish and Not-Selfish by using the attributes present.[03]

Xueying Xu et al. (2020) have studied about the ability of different imputation methods for missing values in mental measurement questionnaires. The authors have explained four different imputations techniques i.e. direct deletion, mode imputation, Hot-deck imputation and Multiple imputation. As per the research observations, they observed that the bias obtained by the Multiple Imputations was the smallest under various missing proportions.[04]

4.Data Introduction

A survey was conducted with 44 question which speaks about an individual's generic demographics, personality traits, health habits, interests and spending habits. Objectives of the survey involves analysing different personalities present in the data and also pandemic's affect on those personalities. Predicting spending kind of a person is also an attempt here. In this research, the aim is not to test a hypothesis about a broader population, but to develop an initial understanding over a small under-researched population which might result in some interesting insights.

The questionnaire was prepared after some research on the topics related to pandemic so that survey would feel open-ended but related. Questionnaire consisted of multiple-choice questions with most of the questions in Likert scale, and every question has sub sections named Before Corona and Present. Survey was designed in such way that individual's identity remains anonymous so that it is filled in sincerely. As the survey was rolled out to the convenient population available and people in contact to that population, 551 responses were recorded in span of 6 days. All the features in the dataset are categorical in nature with 6 demographic features being nominal and remaining 36-38 being ordinal.

4.1.Data Dictionary

Questions/Feature Description	Features	Features
	Before Corona	Present
<i>Generic</i>		
What is your age?	Age (Age groups)	Age (Age groups)
What is your annual income?	Income	Income
What is your employment status?	Emp_stat_Before	Emp_stat_Present
What gender do you identify as?	Gender	Gender
Are you married?	Marital_status	Marital_status
Where is your home located (state) ?	loc	loc
<i>Personality traits</i>		
I take notice of what goes on around me.	Notice_things_Before	Notice_things_Present
I look at things from all different angles before I go ahead.	All_angles_Before	All_angles_Present
I constantly strive to be sincere and productive at work	Sincere_prod_Before	Sincere_prod_Present

I feel lonely in life.	Lonely_Before	Lonely_Present
I worry about my health	Worry_health_Before	Worry_health_Present
I always give to charity.	Charity_Before	Charity_Present
I can quickly adapt to a new environment.	New_env_Before	New_env_Present
I enjoy meeting new people.	Meeting_ppl_Before	Meeting_ppl_Present
I have many different hobbies and interests.	Hob_interests_Before	Hob_interests_Present
I enjoy taking part in surveys.	Surveys_Before	Surveys_Present
How much time do you spend online?	Spent_onli_Before	Spent_onli_Present
I prefer Working from home over going to workspace	WFH_office_Before	WFH_office_Present
Enthusiasm to start something new or have a change in the usual.	Enthu_Before	Ethu_Present
Change in your income due to pandemic	----	Income_Change
<i>Health habits</i>		
Smoking habits	Smoking_hab_Before	Smoking_hab_Present
Drinking habits	Drinking_hab_Before	Drinking_hab_Present
Sleeping habits	Sleeping_hab_Before	Sleeping_hab_Present
I live a very healthy lifestyle.	Healthy_Lifestyle_Before	Healthy_Lifestyle_Present
Which kind of medicine do you prefer	Medi_pref_Before	Medi_pref_Present
<i>Interests</i>		
Interest in Politics.	Pol_interest_Before	Pol_interest_Present
Spending time on Internet.	Internet_interest_Before	Internet_interest_Present
Interest in Economy and Management.	Economy_Manag_intrst_Before	Economy_Manag_intrst_Present
Interest in things related to medicine	Medicine_intrst_Before	Medicine_intrst_Present
Views on Religion	Religion_intrst_Before	Religion_intrst_Present
<i>Spending habits</i>		
I save all the money I can.	Save_all_money_Before	Save_all_money_Present

I prefer branded food products (Organic and Famous food brands) to non branded.	Brand_non-brand_Before	Brand_non-brand_Present
Food preferences	Food_pref_Before	Food_pref_Present
which mode of transport do you prefer?	Mode_of_transport_Before	Mode_of_transport_Present
I keep stock of basic medications	Basic_medications_Before	Basic_medications_Present
I prefer digital content more than going out and spending for entertainment	Digital_content_Before	Digital_content_Present
I am an active investor (Stocks, Mutual funds, Gold e.t.c)	Active_Investor_Before	Active_Investor_Present
I prefer reasonable fee charging educational institutions over expensive institutions	Edu_instit_fee_Before	Edu_instit_fee_Present
How did your spending on data change	— —	change_in_data_consumption
I like spending on gadgets	Spend_on_gadgets_Before	Spend_on_gadgets_Present
I spend on Non-essential/Luxury items also.	Spend_on_Luxury_Before	Spend_on_Luxury_Present
Number of domestic helps	Domestic_help_Before	Domestic_help_Present
I keep track on my household expenses	Track_Household_exp_Before	Track_Household_exp_Present
On a generic scale, what kind of spending person you think you are?	Spending_kind_Before	Spending_kind_Present

4.2.Pre-Processing

4.2.1.Anomaly Detection

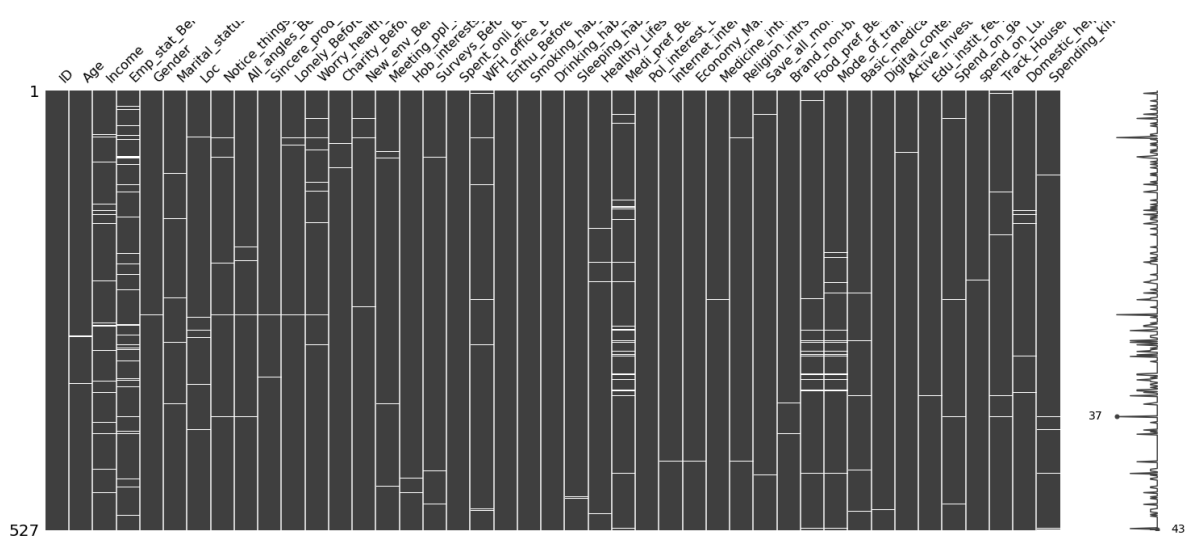
As the data was collected through survey, there was a possibility that data consisted of anomalies or misleading entries. We noticed that there were few responses where responders did not complete the whole survey. Similarly, there could have been a possibility of reluctance in responding to the survey properly, i.e. responding the survey without reading the questions or responding with false information, such as a response where someone responded his location to be Mars. Such responses would have led into misleading inferences and also effected prediction.

Thus, to detect such responses, data was filtered based on an individual's employment status and income. Entries in both the set were filtered and were matched if the same individual have responded in similar

fashion for both the scenarios. 13 responses were noticed to be abnormal with reference to their employment status and income, for example, someone responded themselves to be a student and entered his/her income to be 20+ lakh per annum in both before pandemic and also current situation responses. There were also few responses which seemed fair enough in the employment status and income they have responded with. These 13 filtered responses were then closely studied and 7 entries were finalized as anomaly as a pattern in their responses were noticed. Most of the questions had all answered with same option in both the situations supporting their misleading response of employment status and income. Hence these 7 entries were dropped along with the incomplete responses. This was done such that responses of same individual's were maintained across both the datasets.

4.2.2.Null Value Treatment

After filtering the data from the anomalies and incomplete responses, approximately 2.6% of null values were detected in the data. In Fig 4.2.2.1 and 4.2.2.2, we can observe the spread of null values after anomaly treatment. We can say that the null values are pretty widely spread across the datasets. On right side of the image we can notice the distribution of null values in rows. Another observation is that there are a lot of null values in features employment_stat, Medicine_pref, food_pref and mode_of_transport. Also, the density of null values in these features have increased in the present dataset while the distribution has almost remained same. Even the null values in features in both scenarios mostly don't match which says that individuals have either not responded for before pandemic section or current situation section. Rows which had no spending_kind, were dropped as spending_kind is the target variable in the dataset and imputing target variable null values would not make sense.



Fig(4.2.2.1)

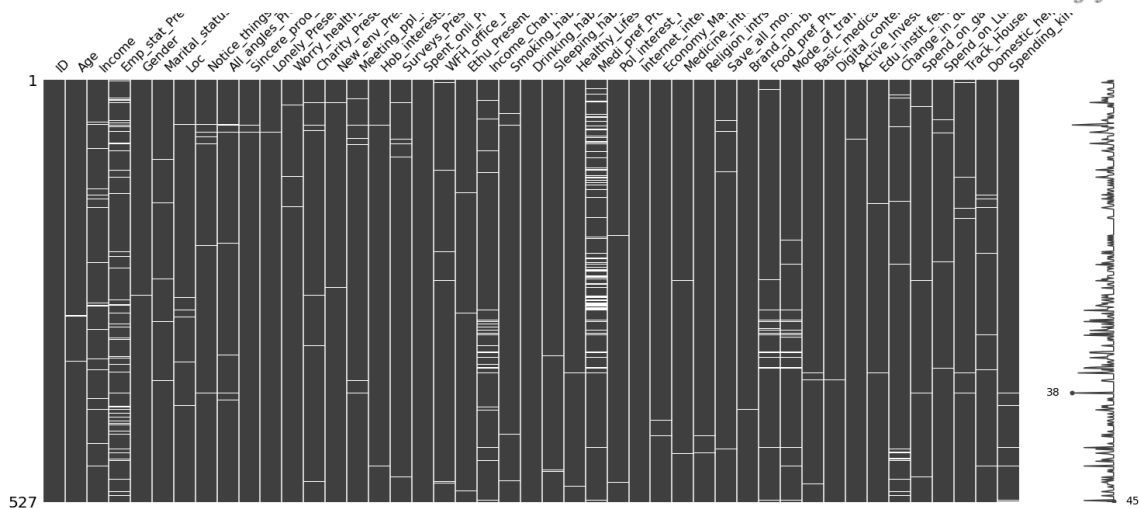


Fig (4.2.2.2)

Understanding the reasons why data are missing is important for handling the remaining data correctly. If values are missing completely at random, the data sample is likely still representative of the population. But if the values are missing systematically, analysis may be biased. To identify the nature of missing values we performed heatmap on nullity correlation matrix of the data. Nullity correlation matrix values ranges from -1 (if one variable appears the other definitely does not) to 0 (variables appearing or not appearing have no effect on one another) to 1 (if one variable appears the other definitely also does). In Fig 4.2.2.3 and Fig 4.2.2.4 we have the heatmap displayed, where most of the values in the matrix are around 0. The features which have values as 1 are due to only one null value present in one of those features which are related. Hence by this we can deduce that the nature of missing values in the data is random.

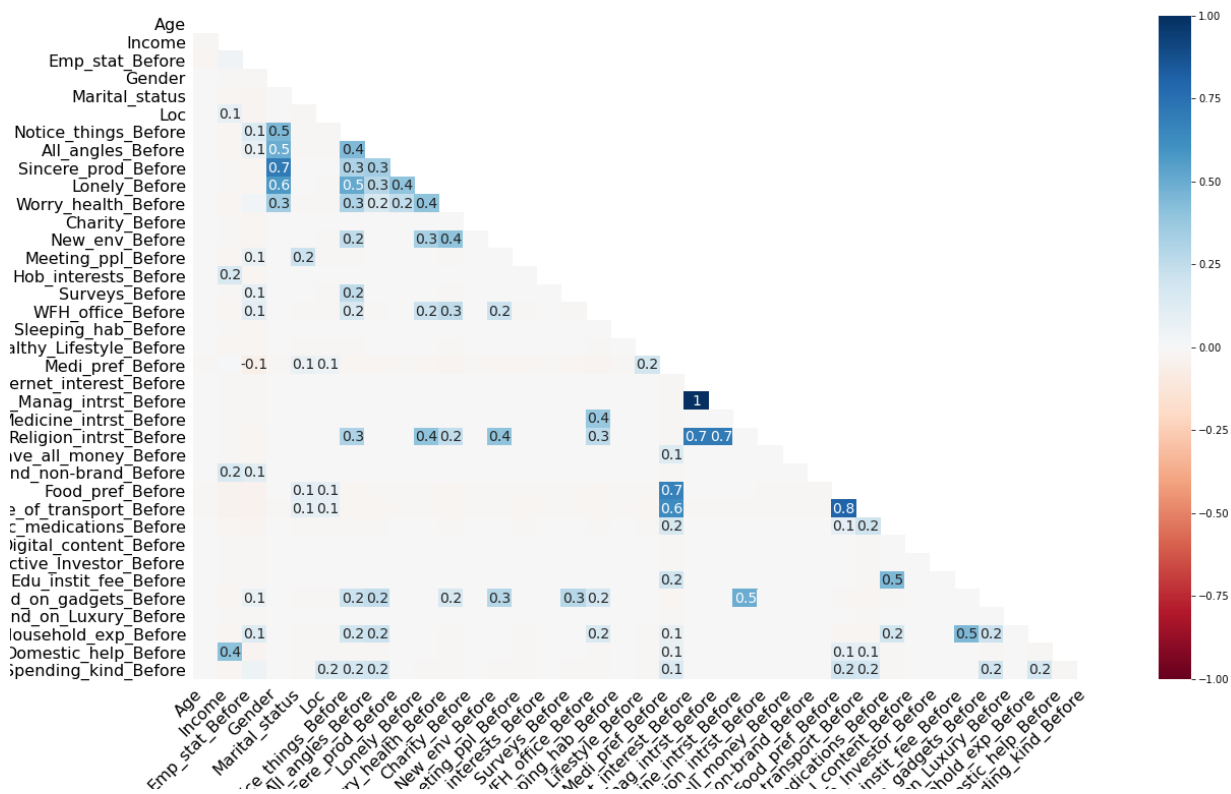


Fig 4.2.2.3

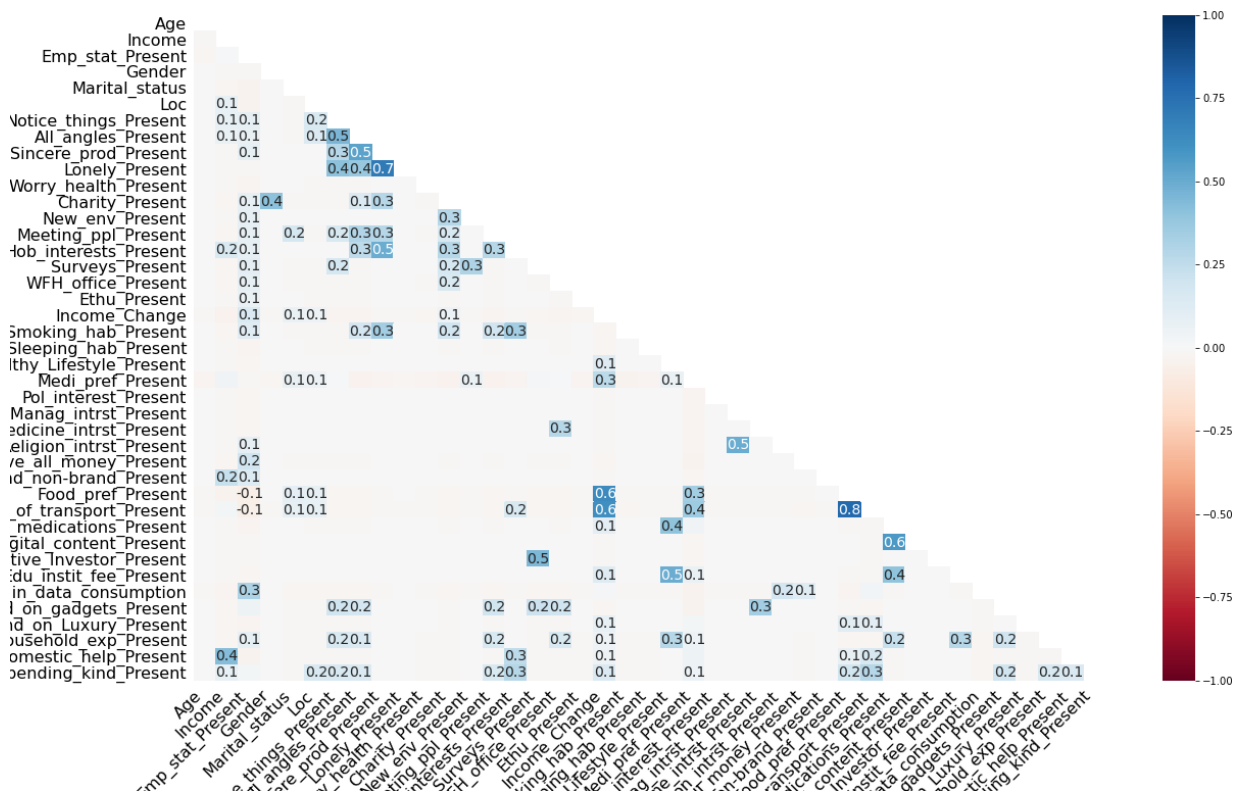


Fig 4.2.2.4

Hence, we wanted to impute these null values in a different way rather than going with the traditional way such that the imputed values would not add bias to the data. After some research on different imputation techniques, we came across a research paper that explains the ability of different imputation methods for missing values in mental measurement questionnaires.

Link: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7045426/>

As per the research, they tested the ability of 4 imputation techniques which were direct deletion, mode imputation, hot-deck imputation, multiple imputation. The biases obtained by multiple imputations are the smallest under various missing proportions. We choose to go with the hot-deck imputation and multiple imputation techniques.

Hot-deck-imputation

Hot-deck imputation with method k-nearest neighbours which uses distance metric to measure the similarity was adopted here. Reason behind using KNN was it measures similarity between the datapoints and imputes the nearest possible value. This method weights samples using the mean squared difference on features for which two rows both have observed data and finds similarity. The only drawback of this package is that it works only on numerical data.

Multiple imputations

Multiple imputation can be used in cases where the data is missing completely at random, missing at random, and even when the data is missing not at random. However, the primary method of multiple imputation is multiple imputation by chained equations (MICE). It is also known as "fully conditional specification" and, "sequential regression multiple imputation." MICE has been shown to work very well on missing at random data which is nearly being exhibited by our null values in our data.

Multivariate imputations via chained equations(MICE) was applied using both r and python modules, to impute the missing values. MICE, by default in r imputes the missing values, by using different algorithms based on feature types. It uses PMM (PREDICTIVE MEAN MATCHING) for numerical data. Logistic Regression for binary data Poly regression for unordered categorical data(where factor>2). Polytomous regression for ordered categorical data where factor>2.[05]

To apply such advance techniques, data needed to be in numerical format, which required encoding the data. As most of the data followed a specific scale, manual encoding was adopted instead of scikit-learn

packages (label encoding) which encodes data in alphabetical order. Below is the scale that we have followed in for encoding:

- Age:
 - 20-25: 0
 - 25-30: 1
 - 30-35: 2
 - 35-40: 3
 - 40-55: 4
 - 55+: 5
- Income:
 - I don't earn right now (maybe in future): 0
 - Under 3 lakh per annum: 1
 - 3 - 5 lakh per annum: 2
 - 5 - 10 lakh per annum: 3
 - 10 - 20 lakh per annum: 4
 - 20+ lakh per annum: 5
- Gender:
 - Male: 0
 - Female: 1
- Marital_status:
 - No: 0
 - Yes: 1
- Loc: Locations entered by the responders were initially generalized into three different geographical topologies and were encoded as below:
 - North India: 0
 - South India: 1
 - Foreign: 2
- For columns on personality traits - (Notice_things_Before, All_angles_Before, Sincere_prod_Before, Lonely_Before, Worry_health_Before, Charity_Before, New_env_Before, Meeting_ppl_Before, Hob_interests_Before, Surveys_Before, Enth_u_Before) and also similar features in present dataset :
 - Strongly disagree: 1

- Disagree: 2
 - Neutral: 3
 - Agree: 4
 - Strongly Agree: 5
- Other columns on spending patterns - (Healthy_Lifestyle_Before, Save_all_money_Before , Active_Investor_Before, Edu_instit_fee_Before, Spend_on_gadgets_Before) and also similar features in present dataset :
 - Strongly disagree: 1
 - Disagree: 2
 - Neutral: 3
 - Agree: 4
 - Strongly Agree: 5
- For columns on Interests - (Pol_interest_Before, Internet_interest_Before , Economy_Manag_intrst_Before, Medicine_intrst_Before , Religion_intrst_Before) and also similar features in present dataset :
 - Not interested: 1
 - Neutral: 2
 - Moderately interested: 3
 - Very interested: 4
- Miscellaneous 1- (WFH_office_Before, Brand_non-brand_Before, Basic_medications_Before, Digital_content_Before, spend_on_Luxury_Before) and also similar features in present dataset :
 - Disagree: 1
 - Neutral: 2
 - Agree: 3
- Spending_kind_Before and Spending_kind_Present:
 - Low spending: 0
 - Medium spending: 1
 - High spending: 2
- Track_Household_exp_Before and Track_Household_exp_Present :
 - No = 0
 - Maybe = 1
 - Yes = 2

- Mode_of_transport_Before and Mode_of_transport_Present :
 - Public (Metro, govt e.t.c) = 0
 - Private (cabs e.t.c) = 1
 - Personal = 2
- Food_pref_Before and Food_pref_Present :
 - Outside (Online & Offline) = 0
 - Home - Cooked = 1
- Medi_pref_Before and Medi_pref_Present:
 - Homemade or Traditional (Ayurvedic e.t.c) = 0
 - English Medicine = 1
- Smoking_hab_Before and Smoking_hab_Present:
 - Never Smoked = 0
 - Former Smoker = 1
 - Tried smoking = 2
 - Current smoker = 3
- Drinking_hab_Before and Drinking_hab_Present:
 - Never = 0
 - Social Drinker = 1
 - Drink a lot = 2
- Sleeping_hab_Before and Sleeping_hab_Present:
 - Early to bed and early to rise = 0
 - Late to bed and early to rise = 1
 - Early to bed and late to rise = 2
 - late to bed and late to rise = 3
- Spent_onli_Before and Spent_onli_Present:
 - No time at all = 0
 - Less than an hour a day = 1
 - Few hours a day = 2
 - Most of the day = 3
- Income_Change, Change_in_data_consumption :
 - Decreased = 0
 - Same = 1

- Increased = 2

Two different datasets were obtained with Hot-Deck (KNN imputer) imputation and MICE imputation techniques applied on the encoded dataset. A different approach apart from these two techniques was adopted where modelling was performed on the data to impute the null values. Initially, dataset was divided into two dataframes containing null values and not null values for a particular feature. The Not null values dataframe was used for training the model and predictions were done on the null values dataframe. Decision tree, Random Forest and Gradient boost were applied where Gradient boost algorithm had better training f1 scores of the three. Below are the f1 scores of feature models with maximum null values in the datasets:

Features	null values	F1 score
Emp_stat_Before	34	0.599016
Worry_health_Before	10	0.359870
Medi_pref_Before	26	0.572356
Food_pref_Before	18	0.564898
Mode_of_transport_Before	20	0.381232

Fig(4.2.2.5)

Features	null values	F1 score
Emp_stat_Present	47	0.608739
Income_Change	24	0.576901
Medi_pref_Present	73	0.580285
Food_pref_Present	16	0.872702
Mode_of_transport_Present	17	0.517939
Change_in_data_consumption	15	0.542101

Fig (4.2.2.6)

On a whole, mean of all the training f1 scores of all the feature models was calculated in order to evaluate performance of this method. Below are the scores:

- mean of all scores: 0.47, max: 0.89, min: 0.21 (for before pandemic responses)
- mean of all scores: 0.50, max: 0.93, min: 0.29 (for current situation responses)

Main objective here was to perform different imputation techniques and have multiple options of cleaned datasets for modelling where performances will be compared and best will be selected.

5.Data Exploration (EDA)

5.1.Relationship between variables

The datasets containing information before the pandemic and present are dominated by categorical variables which are mostly ordinal in type and are arranged into the following segments:

- Demographics
- Personality traits
- Health habits
- Interests
- Spending habits

Since most of the data is categorical and the nature of the analysis is supposed to be change-centric, sectional, groups-based and proportion based, following plots and methods are used to generate a uni-variate, bi-variate and multi-variate analysis:

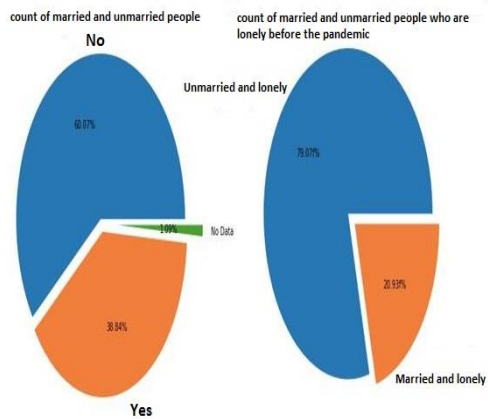
- Count plots with hue
- Barplots
- Pie – Charts
- Stacked bar chart using crosstab
- Masking, slice-indicing, bucketing
- Transition matrices using crosstab
- Chi-square test of independence

The entire analysis has been divided into the following three segments:

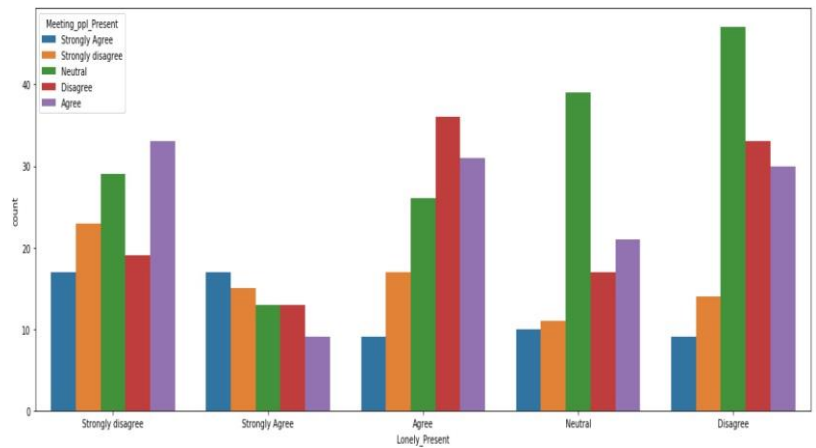
- Behavioural Analysis (analysis on personality traits, habits, interests)
- Economical (analysis on spending habits)
- Combination of behavioural and economical aspects (analysing how people's interests are affecting their spending)

5.1.1. Behavioural Analysis

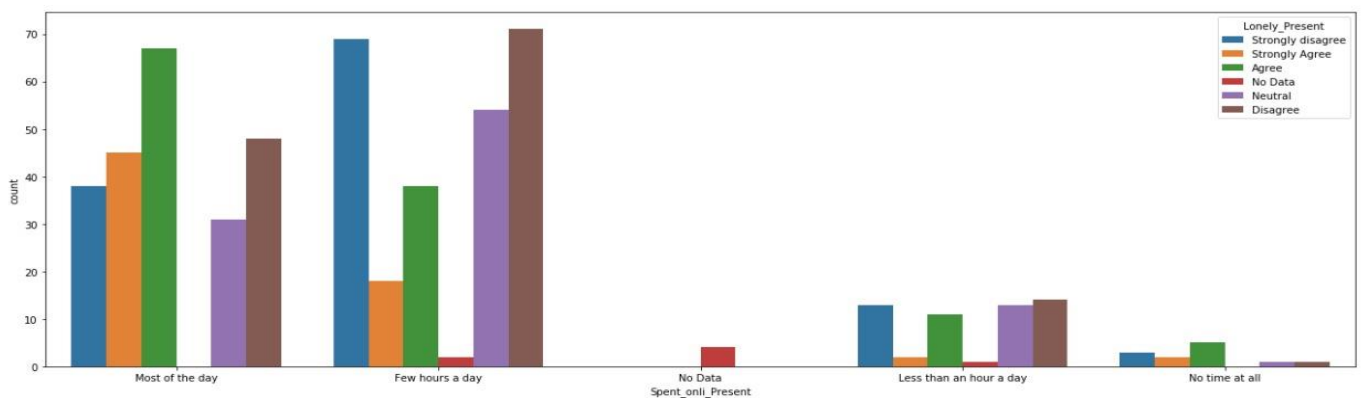
- Analysis on "I feel lonely in life(before and present)"



Fig(5.1.1.1)



Fig(5.1.1.2)



Fig(5.1.1.3)

Fig(5.1.1.1), Fig(5.1.1.2) and Fig(5.1.1.3) show an example of behavioural analysis and explains the following insights below:

- Lonely_Before

Married:

- Married people constitute 20% of the entire number of people who are lonely.
- Married males earning 20+ lakh per annum in the age group of 40-55 years are feeling lonelier, considering their proportion in the total number of people who took the survey, this is a significant observation.

Unmarried:

- Unmarried people constitute 79% of the total number of people feeling lonely before the pandemic.
- This include unemployed youth who does not follow a proper health routine, people in their early 20s earning between 3-5 lakh per annum who have the habit of eating junk and hold irregular sleeping patterns.

- Lonely_Present

- we can observe that overall, the number of people feeling lonely almost got doubled due to the pandemic
- we can also observe that within the people who are strongly agreeing that they are feeling lonely currently, also strongly agreed on the point that they enjoy meeting new people in the current situation as well. From this we can deduce that the restrictions of social distancing, staying away from people, can actually be the potential cause of this loneliness.

➤ Analysis on "I worry about my health" and "I maintain a healthy healthy lifestyle"

Worry_health_Present	Strongly Agree	Agree	Neutral	Disagree	Strongly disagree	Total
Worry_health_Before						
Strongly Agree	34	0	2	0	0	36
Agree	56	69	3	0	1	129
Neutral	43	81	81	3	0	208
Disagree	19	41	7	38	3	108
Strongly disagree	11	13	5	5	18	52
Total	163	204	98	46	22	533

Fig(5.1.1.4)

Healthy_Lifestyle_Present	Strongly Agree	Agree	Neutral	Disagree	Strongly disagree	Total
Healthy_Lifestyle_Before						
Strongly Agree	70	13	10	2	2	97
Agree	30	100	20	22	0	172
Neutral	15	50	120	8	4	197
Disagree	3	12	18	20	3	56
Strongly disagree	1	1	3	1	7	13
Total	119	176	171	53	16	535

Fig(5.1.1.5)

Fig(5.1.1.4) and Fig(5.1.1.5) shows the following insights:

- 49.6% people used to have a healthy lifestyle before which now shifted to 55% .
- 30.4% of the people used to worry about their health which now shifted to 68.8% .
- The huge shift in the people worrying about their health(almost double) and a very negligible shift in the people maintaining a healthy lifestyle tells us an interesting story here:-
 - These can be the people who are actually worried about their health but are not taking any action towards doing something productive and maintaining a healthy routine.
 - If we dig deep into people holding this behaviour, we can observe that they are dominated by unmarried females in the age group of 20-25, who are not earning currently, who spend most

of the day online, and have an irregular sleeping habits of late to bed and late to rise. They are worrying strongly about their health but still not taking any action.

- Earlier, they were neutral about the health aspects of their lives, the pandemic did bring in a change of noticing the importance of health for them, but not to that intensity that they actually start working towards it.
- It's again proved here that end of the day it's one's own determination that can bring in change, not the external situation.
- Having said that, there is an appreciable lot that are actually worried and are taking action by maintaining a healthy routine, these are the males between the age group 20-30 working with an average income of 7 lakh per annum, who had a neutral opinion on health aspects and had a habit of eating outside frequently, which now changed to maintaining a healthy lifestyle, eating home cooked food and waking up early.

➤ Analysis on smoking habits (before and present)

Smoking_hab_Present	Current smoker	Former smoker	Never smoked	Tried smoking	Total
Smoking_hab_Before					
Current smoker	17	15	2	2	36
Former smoker	3	22	12	2	39
Never smoked	2	0	393	7	402
Tried smoking	1	7	33	21	62
Total	23	44	440	32	539

Fig(5.1.1.6)

• NON-SMOKERS :-

- 90% people who didn't have the habit of smoking before the pandemic are still continuing to maintain the same, compared to the proportion of non-smokers the people who have shifted to smoking are very less (7, almost negligible)
- Even though negligible, if we dig a little bit deep and observe the characteristics of these people :-
 - Majority of them, that is 67% of them are in the age group of 35-55 years, earning an average income of 15 lakh per annum, belonging to the states mostly north of India.

Considering the proportion of the respective categories(age, income, location) this is a significant observation.

- 80% of them agreed that they feel lonely in life currently and spend most of the day online ,and also that they worry about their health currently but are still not maintaining a healthy lifestyle.

• **SMOKERS :-**

- Now looking at the people who used to smoke before and are continuing to do so, we can say that 50% of the people who used to smoke before are still continuing and following are their characteristics:-

- 58% of these people are married and are between the age group 40-55 and 55+ and have the irregular sleeping habit of late to bed and early to rise and are also social drinkers.

Considering the proportion of married people and people in the age group of 40-55 and 55+ who have taken this survey ,we can say that this a very significant observation.

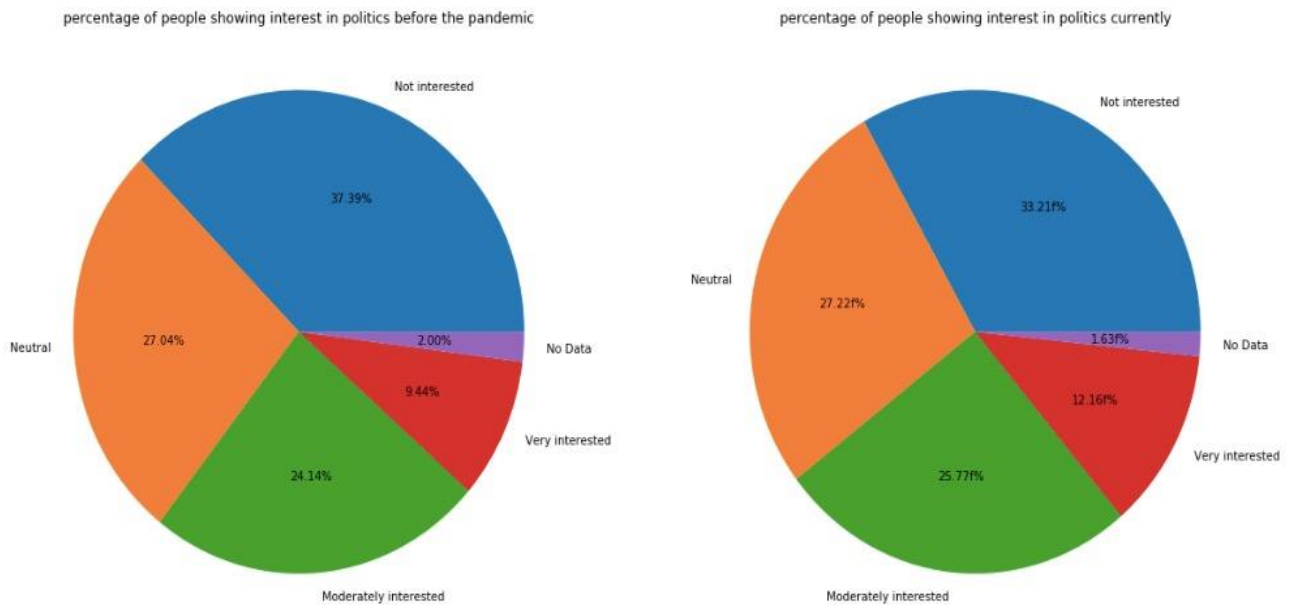
- Now coming to the crucial 50% of the people who actually chose to quit smoking, following are their characteristics :-

- Nearly 70% of these people are youth between 20-30 years of age and have a stable average income of 7 lakh per annum and have also agreed on welcoming certain healthy changes in their lives, like not only worrying about their health but also acting towards it and maintain a healthy lifestyle,
- Nearly 42% of these people agreed that they started being more productive at work and started thinking from different perspectives before taking a decision which wasn't the case before.

➤ **Analysis on "Political Interest"(before and present)**

Pol_interest_Present	Moderately interested	Neutral	No Data	Not interested	Very interested	Total
Pol_interest_Before						
Moderately interested	94	13	2	7	17	133
Neutral	21	117	0	6	5	149
No Data	1	1	7	2	0	11
Not interested	19	17	0	167	3	206
Very interested	7	2	0	1	42	52
Total	142	150	9	183	67	551

Fig(5.1.1.7)



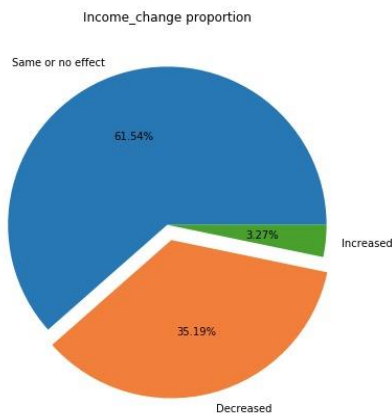
Fig(5.1.1.8)

Refer Fig(5.1.1.7) and Fig(5.1.1.8) for the following insights below:

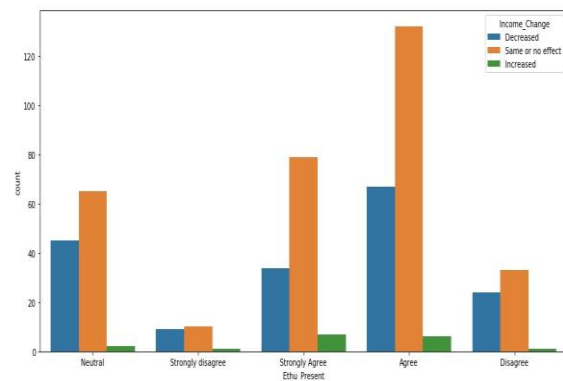
- While majority of the people have shown the same amount of interest towards politics before and currently, we can observe a slight change where a few people who were not at all interested in politics earlier are now showing moderate or high interest in politics, they can be interested towards the governance or the administration, interested in knowing the measures government is taking to curb the situation
- We can also say that they are preferring to be more politically aware to not only abide by the rules that the government is asking to follow but also to know how efficiently the situation is being handled by the people's representatives.
- We dig a little bit deep and understand the specific demographics as of who are the ones who has all of a sudden started showing interest in politics we can see that :-
- It is majorly dominated by employed youth who are in their early 20s ,even though this is a very specific change and also very minimal and has no assurance of sustainability , it can be inferred that the pandemic did actually strike a chord for them to show this drastic change.
- In the real life situation as well we come across people who are volunteering to donate their plasma, and also a category of people who has become more assertive politically and took to twitter to

complain about the situation in Hyderabad to the governor following which a meeting has been called between the governor and the CM.

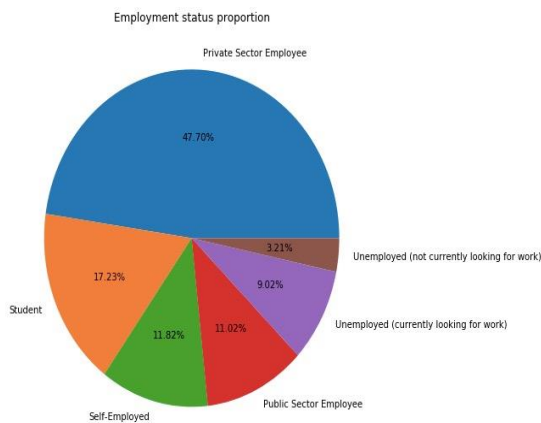
➤ Analysis on "Enthusiasm to start something new"(before and present)



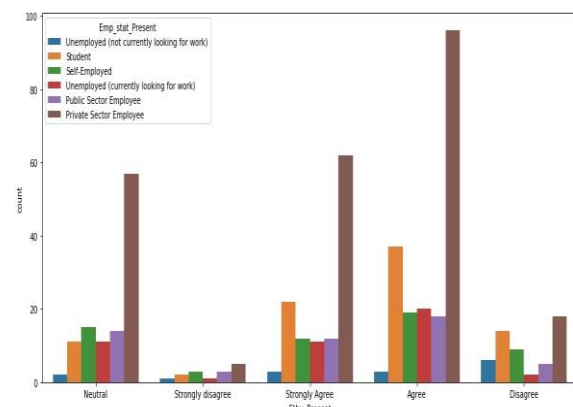
Fig(5.1.1.9)



Fig(5.1.1.10)



Fig(5.1.1.11)



Fig(5.1.1.12)

The figures above show us the following insights:

- The analysis was done by visualising the effect of employment status and income change on the feeling of doing something new, the reason being , we wanted to see whether people whose income levels decreased or people who lost their jobs are the ones chasing to do something new in this direction, or is it not the case.
- Looking at the original proportions of income change and present employment status(in the pie charts) and their proportions in the enthu plots next to them, we observed that the enthusiasm factor is not influenced by one's financial conditions, but might have different perspectives to it,

like starting a new hobby, making a new lifestyle choice, health choice etc.. In this direction a few insights have been derived :-

- 68% of the people who filled the survey were interested in doing something new which now reduced to 60%.
- Even though there is an overall reduction, internally observing the corresponding transition matrices , it was observed that there are incremental shifts where there are people who agreed that they now have the enthusiasm towards doing something new.
- Employed females in the age group 20-30, mostly unmarried are more enthusiastic and all of them strongly agreed on having multiple hobbies currently and are also maintaining a healthy lifestyle.
- They are showing interest towards buying branded food products which are generally costly, all while saving as much money as they can, which justifies their interest in economy management.
- They also have agreed that they are interested in medicinal aspects, which indicates being more health wise.
- It has been observed that people in the age group 30-55 years with a stable income doesn't hold this enthusiasm, and are not preferring new changes in life.

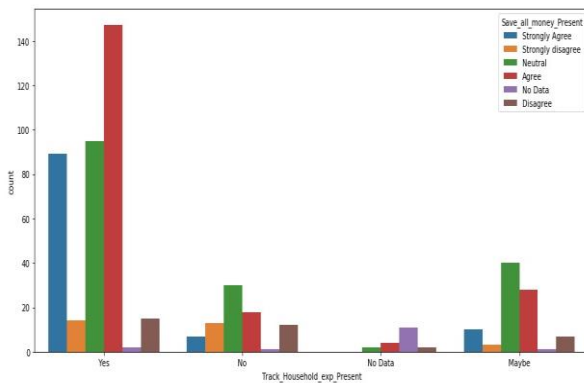
So, the whole thing , in general explains the change in the perspective towards “the way of life”, as a result of their enthusiasm towards making new changes in life.

5.1.2. Economical Analysis

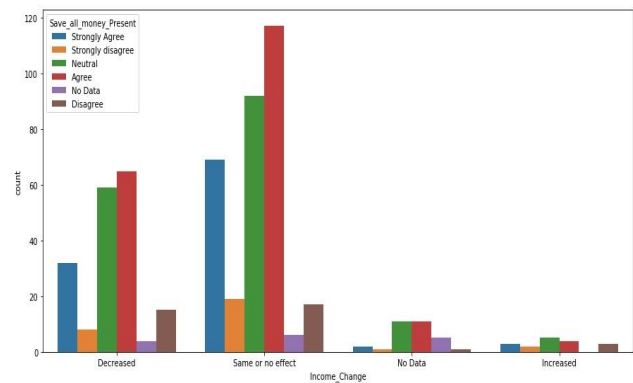
➤ Analysis on "Spending patterns"

Save_all_money_Present	Agree	Disagree	Neutral	Strongly Agree	Strongly disagree	Total
Save_all_money_Before						
Agree	21.455224	0.559701	1.492537	5.223881	0.186567	28.917910
Disagree	5.223881	4.850746	3.731343	1.305970	0.373134	15.485075
Neutral	8.768657	0.559701	24.253731	2.985075	0.932836	37.500000
No Data	0.186567	0.186567	0.559701	0.000000	0.000000	0.932836
Strongly Agree	0.000000	0.186567	0.000000	9.701493	0.186567	10.074627
Strongly disagree	1.119403	0.373134	1.119403	0.559701	3.917910	7.089552
Total	36.753731	6.716418	31.156716	19.776119	5.597015	100.000000

Fig(5.1.2.1)



Fig(5.1.2.2)



Fig(5.1.2.3)

Fig(5.1.2.1), Fig(5.1.2.1), and Fig(5.1.2.1) shows the following economical insights:

- People with decreased income levels has shown a significant changes of shifting from spending on luxury items to not investing on them, from using cabs to opting personal transport, from ordering food online to preferring home cooked food
- Whereas the people whose income didn't get affected neither negatively nor positively still continue to prefer online food and private transport but has cut down spending on luxury items.
- Now there are another interesting category of people whose income increased during the pandemic and are still affording to spend on luxury items and always used to use their personal transport to work
- All these insights are given keeping in mind the dominant personalities of each group(the proportion considered is 62% nearly)
- If we further analyse the spending group of people whose income increased and are spending almost in the same way as before, the following things has been observed :-
 - Although on an average the spending kind showed that they turned towards medium spending, internally it has been observed there are significant high spending people as well (all of them are unmarried),and 70% of this category agreed that they still spend on luxury.
 - All of them are private sector employees and has received a salary rise while being in the same sector, most of them belong to the income group 5-10 lakh per annum between age 25-30 years, this can be due to job switches or hikes.
- Every Age group has decreased their expenditure (So, Irrespective of Age group every individual decreased their cash inflow into the system --> Decreases Liquidity).

- Currently, W.r.t people under high Spending and Medium spending categories, the gender based difference in spending is not that significant, but the difference is significant w.r.t low spending category. It was observed that females are spending less when compared to males. Reasons can be restrictions on shopping, parlour expenses or are probably saving for a better future.
- In the Present situation, Low spending people are showing more interest in Economy_management than before.
- Surprisingly only 10% of ppl from each other groups changed to organic foods/Branded food products
 - People who used to be neutral towards maintaining a healthy lifestyle have now changed to maintain a healthy lifestyle and are preferring branded food products over non-branded ones
- Usually due to Work from Home, Online Entertainment, Online Education, Kids Management the Data Consumption will increase, but 10% decreased their Consumption.
- People who are Self-Employed majorly (~ 60%) decreased their data consumption, this can be due to Lose of business/clients currently (Extended time period for projects which in turn decreases their income).

5.2. Multi-Collinearity

The Multiple correspondence analysis (MCA) is used for summarizing and visualizing a data table containing more than two categorical variables[06]. It can also be seen as a generalization of principal component analysis when the variables to be analyzed are categorical instead of quantitative. MCA is generally used to analyse a data set from survey. The goal is to identify:

- A group of individuals with similar profile in their answers to the questions
- The associations between variable categories

The datasets do exhibit multi-collinearity to some extent, where in there are a few features that give information that has already been captured by other features.

For a large multi-variate categorical data, there are specialised statistical techniques dedicated to categorical data analysis, such as simple and multiple correspondence analysis (MCA). These methods make it possible to analyse and visualise the association (i.e., correlation) between a large number of

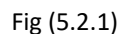


Fig (5.2.1) shows an example of very less multi-collinearity in the data by visualizing before pandemic responses.

The categorical data distribution has been observed through bar graphs and the skewness has been noted.

Outliers are extreme values that we come across, where they may be influential to the model or not. With respect to our data, outliers are those misleading/random answers which are not sensible(anomalies). In this multi variate dataset, no such extreme values were found.

The variable significance:

- With respect to capturing a significant change (between before pandemic and current situations) has been noted by performing **Wilcoxon signed ranked test** and it has been observed that there are 27 such columns which are significant.
- This test is used when the samples are related or matched in some way or represent two measurements of the same technique. More specifically, each sample is independent, but comes from the same population.
 - H0: No significant difference in responses by individuals for a question before and during pandemic
 - HA: significant difference in responses by individuals for a question before and during pandemic
- With respect to target column, has been noted by performing chi-square test of independence.

Features	Pvalues
Emp_stat	0.001774
Notice_things	0.000000
All_angles	0.000000
Sincere_prod	0.049152
Lonely	0.000000
Worry_health	0.000000
Charity	0.030867
New_env	0.018271
Meeting_ppl	0.000000
Hob_interests	0.553008
Surveys	0.085317
Spent_onli	0.000000
WFH_office	0.000000
Enthu	0.011001
Smoking_hab	0.039912
Drinking_hab	0.000012
Sleeping_hab	0.000000
Healthy_Lifestyle	0.085638
Medi_pref	0.000010
Pol_interest	0.045934
Internet_interest	0.000000
Economy_Manag_intrst	0.000562
Medicine_intrst	0.000000
Religion_intrst	0.111576
Save_all_money	0.000000
Brand_non-brand	0.266819
Food_pref	0.000000
Mode_of_transport	0.000000
Basic_medications	0.000000
Digital_content	0.000000
Active_Investor	0.386860
Edu_instit_fee	0.012253
Spend_on_gadgets	0.141569
spend_on_Luxury	0.000000
Track_Household_exp	0.000000
Spending_kind	0.000000

Fig (5.5.1)

Features
Emp_stat
Notice_things
All_angles
Sincere_prod
Lonely
Worry_health
Charity
New_env
Meeting_ppl
Spent_onli
WFH_office
Enthu
Drinking_hab
Sleeping_hab
Medi_pref
Internet_interest
Economy_Manag_intrst
Medicine_intrst
Save_all_money
Food_pref
Mode_of_transport
Basic_medications
Digital_content
Edu_instit_fee
spend_on_Luxury
Track_Household_exp
Spending_kind

Fig (5.5.2)

Fig (5.5.1) shows the features and their p-values that define significant difference in their distributions between before pandemic responses and current situation responses. Hobbies and interests has the

largest p-value stating that is no significant change in the responses received in that aspect. Also, we can observe that there are few other features with high p-value (>0.05) which tells us that there is no significant change in population's views in those aspects.

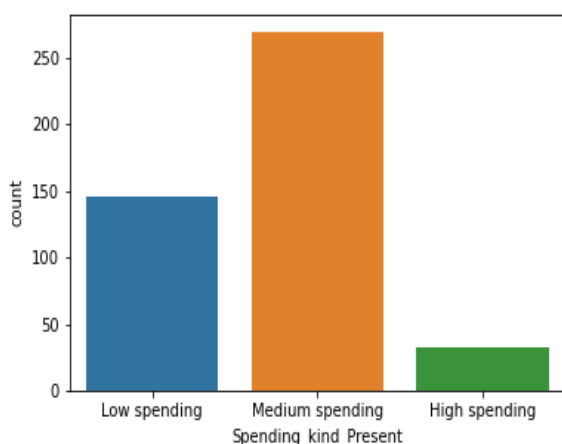
Fig (5.5.2) gives us the features which have significant change in their responses between before pandemic and current situation.

5.6.Class Imbalance and its treatment

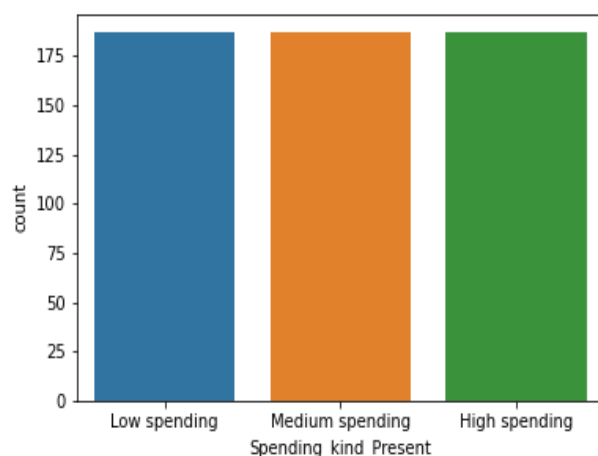
If the datasets intended for classification problems like Sentiment Analysis, Medical Imaging or other problems related to Discrete Predictive Analytics (for example-Flight Delay Prediction) have an unequal number of instances (samples or data points) for different classes, then those datasets are said to be imbalanced.[Refer Fig(5.6.1)]

There are a number of methods available to oversample a dataset used in a typical classification problem. The most common technique is known as SMOTE: Synthetic Minority Over-sampling Technique. To create a synthetic data point, take the vector between one of those k neighbours, and the current data point. Multiply this vector by a random number x which lies between 0, and 1. Add this to the current data point to create the new, synthetic data point.

The dependent variable (spending kind) is a multi class variable with 3 classes (high, medium, low) and was highly imbalanced with the classes in the ratio, and was treated by using smote. After applying smote the count plot can be seen below.(Refer Fig(5.6.2))

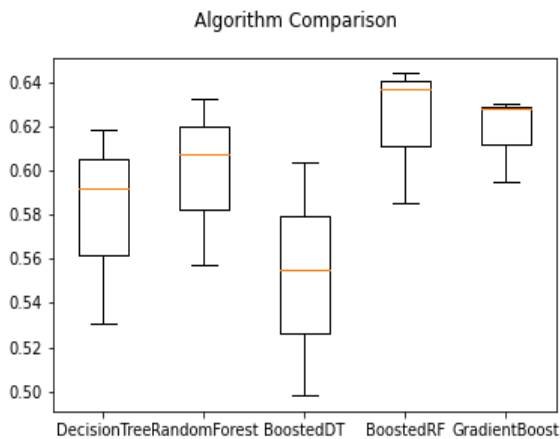


Fig(5.6.1)

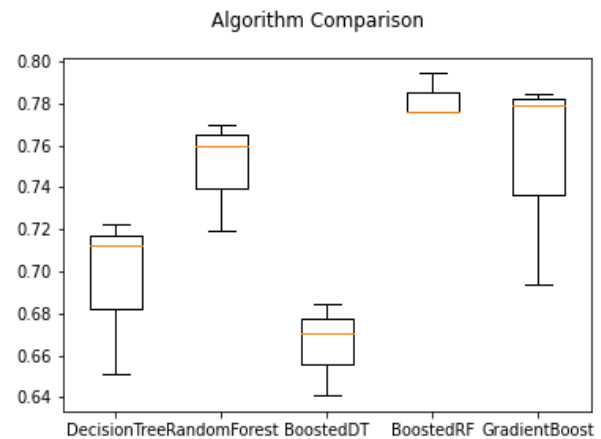


Fig(5.6.2)

Here are the results of modelling we have performed on balanced and imbalance datasets:



Fig(5.6.3)



Fig(5.6.4)

Fig(5.6.3) represents the box plot comparisons of different algorithm without smote and Fig(5.6.4) represents the box plot comparisons of different algorithm on treated dataset(SMOTE)

We can notice the consistency of the balanced dataset is better than imbalanced datasets, but the models with balanced datasets were not able to predict all the classes. Gradient boost algorithm was performing well with imbalance datasets with predicting all the classes fairly better than other models.

6.Feature Engineering

6.1.Scaling

The main idea behind normalization/standardization is always the same. Variables that are measured at different scales do not contribute equally to the model fitting & model learned function and might end up creating a bias. Thus, to deal with this potential problem feature-wise normalization scaling is required used prior to model fitting.

In our case, the dataset consists of ordinal and nominal features and to make all the features of equal importance, scaling was required to use the data in distance metric based models. Also, scaling the data is a prerequisite for Clustering Problems. Hence Min-Max Scaler from scikit-learn Library was applied to the data.

Min-Max scaler is a normalization technique that transforms the features in the range of 0 to 1.

6.2.Feature Selection

before_features	
	Income
	Gender
	Notice_things_Before
	All_angles_Before
	Lonely_Before
	Charity_Before
	Meeting_ppl_Before
	Spent_onli_Before
	WFH_office_Before
	Smoking_hab_Before
	Drinking_hab_Before
	Healthy_Lifestyle_Before
	Religion_intrst_Before
	Save_all_money_Before
	Brand_non-brand_Before
	Food_pref_Before
	Basic_medications_Before
	Digital_content_Before
	Edu_instit_fee_Before
	Spend_on_gadgets_Before
	spend_on_Luxury_Before
	Track_Household_exp_Before
	Domestic_help_Before
	Spending_kind_Before

Fig (6.2.1)

present_features	
	Gender
	All_angles_Present
	Hob_interests_Present
	Income_Change
	Smoking_hab_Present
	Economy_Manag_intrst_Present
	Save_all_money_Present
	Brand_non-brand_Present
	Food_pref_Present
	Basic_medications_Present
	Digital_content_Present
	Active_Investor_Present
	Edu_instit_fee_Present
	Spend_on_gadgets_Present
	Spend_on_Luxury_Present
	Track_Household_exp_Present
	Spending_kind_Present

Fig (6.2.2)

Statistical methods were used to derive relationship between the features. Two types of statistical test were performed which are:

- Test of difference in proportion
- Test of independence

For test of difference in proportion Wilcoxon signed ranked test was performed and for test of independence **Chi-Square test of independence** was performed with below hypothesis.

- H0: No significant relationship between the Variables. The variables are independent.
- HA: A relationship between the variables exists.

The Chi-Square test of independence is used to determine if there is a significant relationship between two categorical variables. Hence, the test was performed to derive features which have significant relationship with the target variable.

There were 23 features in before pandemic responses which had significant relationship with its target variable ('Spending_kind_before') and 16 features in the current situation responses which had significant relationship with its target variable ('Spending_kind_Present').

Fig (6.2.1) shows us the features which exhibit significant relationship with the target variable in before pandemic responses.

Fig (6.2.2) shows us the features which exhibit significant relationship with the target variable in current situation responses.

6.3.Dimensionality reduction

MCA i.e. Multiple Correspondence Analysis is also known to be the counter part of Principal component analysis for categorical features. Its takes indicator matrix as input and associations between variables are uncovered by calculating the chi-square distance between different categories of the variables and between the individuals (or respondents). These associations are then represented graphically as "maps", which eases the interpretation of the structures in the data. Oppositions between rows and columns are then maximized, in order to uncover the underlying dimensions best able to describe the central oppositions in the data. As in factor analysis or principal component analysis, the first axis is the most important dimension, the second axis the second most important, and so on.

Since the dataset has very less multi-collinearity , the principal components derived were capturing were less variance i.e. the first dimension explained around 2.86% of variance. So, the components derived were from MCA were not used.

7. Supervised Learning(Classification)

Objective: Dealing this as a prediction problem(classification) and predicting what spending category(High/Medium/Low), a person with a set of characteristics in the present situation will belong to.

7.1. Base Model

As the data present consisted of all categorical features which exhibit very less correlation and show very less linearity, a decision-based model such as Decision Tree was choose to be the base model. Main objective in the predictive model is to predict spending kind based of an individual based on all the responses entered.

The general motive of using Decision Tree is to create a training model which can use to predict class or value of target variables by learning decision rules inferred from prior data (training data). Decision tress often mimic the human level thinking so it's simple to understand the data and make some good interpretations.

Also, Decision tree requires very less effort for data preparation during pre-processing and also does not require any normalization or scaling of data. It is also very intuitive and easy to explain. Although, for a Decision tree sometimes calculation can go far more complex compared to other algorithms and often requires higher time to train the model.

	precision	recall	f1-score	support
0	0.25	0.43	0.32	14
1	0.90	0.77	0.83	102
2	0.48	0.58	0.52	19
accuracy			0.71	135
macro avg	0.54	0.59	0.56	135
weighted avg	0.77	0.71	0.73	135

```
-----
[[ 6  7  1]
 [12 79 11]
 [ 6  2 11]]
```

Fig (7.1.1)

	precision	recall	f1-score	support
0	0.43	0.48	0.45	42
1	0.70	0.64	0.67	83
2	0.31	0.40	0.35	10
accuracy			0.57	135
macro avg	0.48	0.50	0.49	135
weighted avg	0.59	0.57	0.58	135

```
-----
[[20 19  3]
 [24 53  6]
 [ 2  4  4]]
```

Fig (7.1.2)

Fig (7.1.1) and Fig(7.1.2) show us the testing accuracy of base model (Decision Tree) after trying to predict Spending_kind_before from before pandemic responses and Spending_kind_Present from current situation responses. Training accuracy of both models was found around 60% and 52%.

7.2. Other Models

7.2.1 K Nearest Neighbour

kNN stands for k Nearest Neighbour. In data mining and predictive modelling, it refers to a memory-based (or instance-based) algorithm for classification and regression problems. In classification problems, the label of potential objects is determined by the labels of closest training data points in the feature space. The determination process is either through "majority voting" or "averaging". In "majority voting", the label of object is assigned to be the label which most frequent among the k closest training examples. In "averaging", the object is not assigned a label, but instead, the ratio of each class among the k closest training data points.

KNN model was built on the dataset and an accuracy of 54% was achieved.

7.2.2 Random Forest

It is an ensemble tree-based learning algorithm. The Random Forest Classifier is a set of decision trees from randomly selected subset of training set. It aggregates the votes from different decision trees to decide the final class of the test object.

Accuracy of 59% was achieved with Random forest.

	precision	recall	f1-score	support
0	0.48	0.31	0.38	39
1	0.67	0.86	0.75	85
2	1.00	0.09	0.17	11
accuracy			0.64	135
macro avg	0.72	0.42	0.43	135
weighted avg	0.64	0.64	0.60	135


```

-----
[[12 27  0]
 [12 73  0]
 [ 1  9 11]]

```

Fig(7.2.2.1)

7.3 Final Model(Gradient Boost)

Gradient boosting is a type of machine learning boosting. It relies on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error. The key idea is to set the target outcomes for this next model in order to minimize the error.

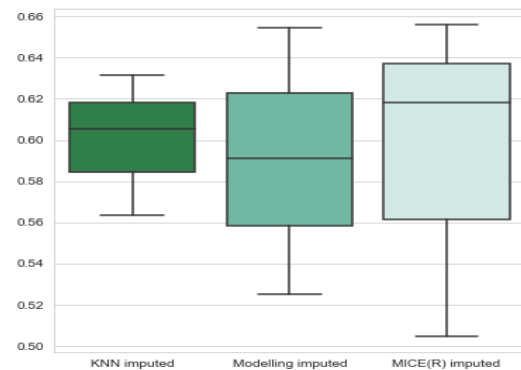
In our data, it gave us the test accuracy of 66% and train accuracy of 60%. It able to classify the "high spending" category comparatively better than the other models. This can be due to the fact that gradient

boost learns from the past errors. Therefore, the influence of imbalanced data is the least. (Training accuracy:60%)

	precision	recall	f1-score	support
0	0.60	0.48	0.53	44
1	0.70	0.83	0.76	81
2	0.75	0.30	0.43	10
accuracy			0.67	135
macro avg	0.68	0.53	0.57	135
weighted avg	0.67	0.67	0.66	135

```
[[21 23 0]
 [13 67 1]
 [ 1  6 3]]
```

Fig(7.3.1)



Fig(7.3.2)

	precision	recall	f1-score	support
0	0.44	0.57	0.49	42
1	0.67	0.59	0.62	82
2	0.12	0.09	0.11	11
accuracy			0.54	135
macro avg	0.41	0.42	0.41	135
weighted avg	0.55	0.54	0.54	135

```
[[24 16 2]
 [29 48 5]
 [ 2  8 1]]
```

Fig(7.3.3)

Fig(7.3.1) represents the classification report of gradient boost algorithm on imbalanced dataset and Fig(7.3.2) represents the classification report of gradient boost algorithm on balanced dataset.

Fig(7.3.3) represents the gradient boost algorithm's performance on differently imputed dataset

From all the classification reports (DT, RF, KNN, GB), it has been observed that the scores for the “high spending” category are least indicating that people are not spending much and are tending to save money.

8.Unsupervised Learning (Clustering)

objective: To observe dominant personalities (with respect to interests, preferences and spending patterns) before COVID and how did it change in the present situation, we used the following clustering techniques:

- K-Means
- Hierarchical(Agglomerative) Clustering
- K-Medoids
- K-Modes

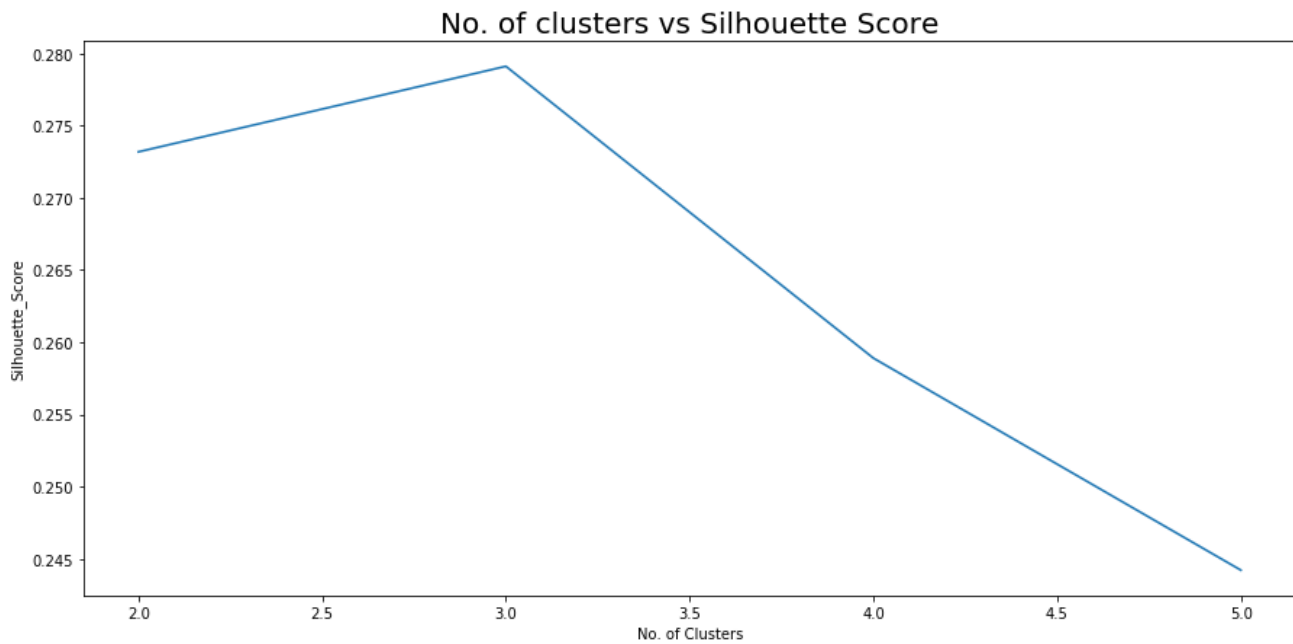
8.1.K-Means

The objective of K-means is simple: group similar data points together and discover underlying patterns. To achieve this objective, K-means looks for a fixed number (k) of clusters in a dataset. In other words, the K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.

K-means clustering is done on variables that we consider important and indicative of the individuals personality and tried a combination of them with varying parameters(MAX_ITER,N_INIT).

The following results were observed with the k means clustering:

- Before dataset: Maximum silhouette of 0.28 with three clusters, with an inertia of 1515.73 has been observed for the before situation.(Refer Fig(8.1.1))
- Present dataset: Silhouette score of 0.26 with five clusters, with an inertia of 1055 were seen.



Fig(8.1.1)

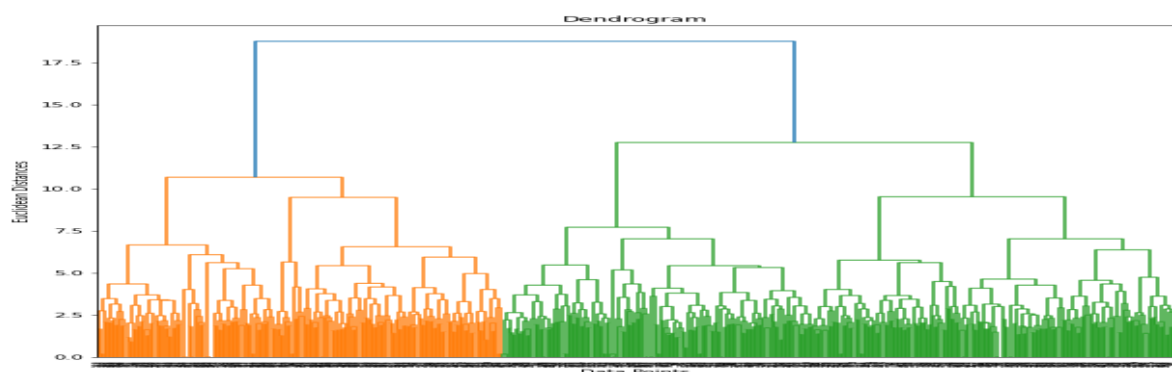
8.2 Hierarchical(Agglomerative) Clustering

Agglomerative Hierarchical clustering Technique: In this technique, initially each data point is considered as an individual cluster. It uses a bottom-up approach, wherein each data point starts in its own cluster. These clusters are then joined greedily, by taking the two most similar clusters together and then merging them.

The following results were observed with the hierarchical clustering:

- Before dataset: Maximum silhouette of 0.266 with three clusters, were observed.
- Present dataset: Silhouette score of 0.28 with five clusters, were observed.

These scores are obtained on those selected features as that of kmeans.



Fig(8.2.1)

8.3 K-Medoids

The idea of K-Medoids clustering is to make the final centroids as actual data-points. This result to make the centroids interpretable. Also, this is an efficient algorithm in clustering in order to cluster categorical data.

The cost in K-Medoids algorithm is given as

$$c = \sum_{C_i} \sum_{P_i \in C_i} |P_i - C_i|$$

Fig(8.3.1)

The K-Medoids clustering when applied to the multivariate categorical dataset, did not give promising results :

- Before dataset: The optimum number of clusters could not be identified from the elbow plot and also, overlapping clusters were formed.
- Present dataset: Two to four clusters with same inertia were formed where a few clusters were empty.

8.4 K-Modes

k-modes provides a much-needed alternative to k-means when the data at hand are categorical rather than numeric.

The k-modes algorithm tries to minimize the sum of within-cluster Hamming distance from the mode of that cluster, summed over all clusters. If a dataset has m categorical attributes, the mode vector Z consists of m categorical values, each being the mode of an attribute. The distance metric used for K-modes is instead the Hamming distance from information theory. The Hamming distance (or dissimilarity) between two rows is simply the number of columns where the two rows differ.[08]

Present Cluster	1	2	3	4	Total
Before Cluster					
1	103	43	35	35	216
2	52	30	16	26	124
3	33	16	14	14	77
4	49	27	10	18	104
Total	237	116	75	93	521

Fig(8.4.1)

For both the datasets, four clusters with costs of 22 were observed.

Although, these clusters were not that distinct(overlapping), a few dominant clusters were observed(Refer Fig(8.4.1))

In the before and present situation(Refer Fig(8.4.1))

- cluster 1 is the dominant cluster which includes people having lenient behaviour.
- cluster 2 includes the confused people (Most of them are neutral in terms of their responses)
- cluster 3 includes smart people who hold strong political opinion and are balanced with respect to their spending behaviour.
- cluster 4 includes lazy and irresponsible people who do not show interest on important aspects of life and are irresponsible when it comes to spending.(spending on luxury items, non-essentials etc)

9.Web Application

Data scientists mould data in various ways to unearth insights from it. And since these outcomes are used for decision-making, it is paramount for data scientists to write production-level code well. However, sometimes, the practices data scientists implement are tedious to utilize in production.

Data scientists often write code to evaluate the data by exploratory data analysis, check several data points, and outliers. These algorithms are only good for getting an idea about how data is spread, but are mostly purposeless in production.

Besides, the algorithms come back from the production team to data scientists for making changes in them to add features or optimize the code for improved efficiency. Such practices are cumbersome and ineffective in this highly competitive marketplace.

Now, with an app framework called Streamlit, data scientists can build machine learning tools right from the beginning of the project. Utilizing Streamlit, data scientists can visualize their code output while analyzing data.

In Our Project using Streamlit, we created a web app which showcases the work we have done, from the background research to evolving to the problem statement with a detailed exploratory data analysis and also models are deployed dynamically and results were observed.

Link for the web app: <http://52.15.116.243:8501>

10.Results

- The classification objective of predicting the spending category(High/Med/Low) of a person with a set of characteristics in the present situation has been achieved by using gradient boost model(best performance).
- The training accuracy of 60% and the testing accuracy of 66% was achieved by using gradient boost.
- The gradient boost performance on the differently imputed dataset(KNN,MICE,Modelling) ,have been visualised using box plot and the algorithm performed best on the mice imputed dataset.[Refer Fig(7.3.2)]
- From all the classification reports (Decision tree ,Random Forest, Gradient Boost), it has been observed that the scores for the "high spending" category are least indicating that in the present situation people are not willing to spend much and are tending to save money.
- Out of all the clustering techniques applied(K-Means, Agglomerative ,K-Modes), K-Modes comparatively gave better results whereas K-Means and agglomerative gave overlapping clusters.
- K-Modes gave four clusters with a cost of 22 for both the before and present dataset.
- Cluster 1 continued to be the dominant cluster containing people showing lenient behaviour who are mostly neutral but specific about few things.

11.Conclusion

The gradient boost model gave optimal accuracy when it comes to categorizing people in to high/medium/low spending category but the association of the category to which they belong are influenced by individuals thinking and are not clearly defined. It has been clearly established that the pandemic made people cut down their expenditure and save money.

Most favourable results were not obtained as distinct clusters indicating dominant personalities (overlapping clusters) were not formed. The possible reason can be the that the personalities are so different (w.r.t. interests, hobbies and preferences) that they could not be densely clustered into one particular cluster. Even though overlapping clusters were formed but the traces of patterns found are a positive sign indicating the feasibility of clustering technique to the field of behavioural analysis.

12. Drawbacks

- Behavioural and economical analysis done is subject to time.
- Questionnaire preparation could have been more thoughtful
- Respondent's transparency
- Data collected is subject to regional, age and occupational bias

13. Vision

- Potential questionnaires with the help of domain experts
- Ideal techniques to deal with a multi variate categorical dataset(Mice, Fisher's Test, MCA etc)
- Optimised Data collection (Sampling techniques- Cluster sampling, Stratified sampling etc)
- NLP or advanced Concepts for sentiment analysis
- To achieve model interpretability and derive detailed inferences, Multinomial logistic regression can be used.
- Since our features are interactive, we can use feature engineering to generate higher order features and observe modelling results.

14. Applications

- Educational Institutions: Analyse the change in student's behaviour in order to improve their performances and accordingly suggest reforms in the education system.
- Private Sector: Analyse employee's behaviour and make changes accordingly to improve the productivity.
- Analysing criminal mindset: Analysing criminal mindsets to understand the reason behind their abnormal/unusual behaviour and help build a civilised society.
- This analysis can be extended to respective industries: Based on individuals consumptions and spending patterns, the respective industrial growth can be noticed.(digital content, pharmacy industry, online transport, online food etc)

15. Executive Summary

- Problem statement - To analyse the effect of COVID-19 on Individuals' personality traits and their spending patterns
 - Objectives :-
 - To add social value by bringing to surface different mental conditions and the effect of the pandemic on the same.
 - To add commercial value, by extrapolating the results of Individual consumption and spending patterns to observe and predict industrial growth(pharma, online entertainment, transportation, food etc..).
 - Effect on Individual personality traits :-
 - Irrespective of one's lifestyle(health habits) , work life balance, it has been observed that people are still feeling lonely.
 - It was observed that, interestingly age groups 20-25 and 40-55 and above are spending more time online watching digital content, the reason can be absolutely no responsibilities or reduction in the responsibilities in both the age groups and having more time than ever to spend, people who are in their mid 30s groups have given very less priority to spending time online and watching digital content as they are surrounded with more responsibilities and less time to spend.
 - People who actually chose to quit smoking, have following characteristics :-
 - Nearly 70% of these people are youth between 20-30 years of age and have a stable average income of 7 lakh per annum and have also agreed on welcoming certain healthy changes in their lives, like not only worrying about their health but also acting towards it and maintain a healthy lifestyle,
 - Nearly 42% of these people agreed that they started being more productive at work and started thinking from different perspectives before taking a decision which wasn't the case before.
 - 30.4% of the people used to worry about their health which now shifted to 68.8% .
- The number of people trying to save the money currently almost got doubled if not exactly doubled. It has also been observed that people with both reduced and same income levels are showing interest in saving currently and are keeping track of household expenses to facilitate this process.

➤ Effect on Individuals' spending :-

- People with decreased income levels has shown a significant changes of shifting from spending on luxury items to not investing on them, from using cabs to opting personal transport, from ordering food online to preferring home cooked food
- Now there are another interesting category of people whose income increased during the pandemic and are still affording to spend on luxury items and always used to use their personal transport to work
- Every Age group has decreased their expenditure, resulting to the decreased cash flow into the system therefore decreasing the Liquidity.
- It was observed that females are spending less when compared to males. Reasons can be restrictions on shopping, parlour expenses or are probably saving for a better future.
- It was observed that females are spending less when compared to males. Reasons can be restrictions on shopping, parlour expenses or are probably saving for a better future.
- In the Present situation, Low spending people are showing more interest in Economy management than before.

➤ Future Enhancements :-

The use of psychometric analysis is not that prominent in the industry today, we therefore can extend this to different industries(entertainment,private sector, medicine etc..) and educational institutions and prepare useful questionnaires and derive insights that will be socially and commercially informative.

16. References

- [01] <https://en.wikipedia.org/wiki/Psychometrics>

- [02] Charu Nath, Rohit Kumar Akhairamka, Sanchit Bhatia and Varsha Ahuja, 2011, Behavioural Analysis Using Data Clustering: BIOINFO Computational Optimization, Volume 1, issue 1, 2011.

- [03] Akexandros Ladas, Uwe Aickelin, Jon Garibaldi, 2013, Using Clustering To Extract Personality Information from Socio Economic Data: SSRN Electronic Journal, Volume1, Page 1-8, 2013.

- [04] Xueying Xu, Leizhen Xia, Qimeng Zhang, Shaoning Wu, Mingcheng Wu, and Hongbo Liu, 2020, The ability of different imputation methods for missing values in mental measurement questionnaires: BMC Med Res Methodol, volume 20, PMCID: PMC7045426.

- [05] <https://datascienceplus.com/imputing-missing-data-with-r-mice-package/>

- [06] https://en.wikipedia.org/wiki/Multiple_correspondence_analysis

- [07] <http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/114-mca-multiple-correspondence-analysis-in-r-essentials/>

- [08] <https://medium.com/@davidmasse8/unsupervised-learning-for-categorical-data-dd7e497033ae>