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

Interim Report

Submitted To

GREAT LEARNING, HYDERABAD

For the Capstone Project

By

ANIRUDH REDDY ITIKELA

JAGTHAP ROHAN KUMAR

KAKARLA SRI RANGANADH

MEGHANA ANANTANENI

SHREYASHI TIWARI

Capstone Project–Interim 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/1MeblsDp1F5Ryq1-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.

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.

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.

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 informations. 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.

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.

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*
Generic		
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
Personality traits		
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
Health habits		
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
Interests		
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
Spending habits		
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:

- Misleading responses or Anomalies in the data were identified by closely looking into the data based on individual's employment status and income.
- 7 entries were identified to be misleading with a pattern in their responses (like filling up every question with same response etc.) ,supporting the misleading entries in the employment status and income.
- Approximately 2.6% of null values were detected in the data.
- Incomplete responses were also identified and removed in a way which would maintain the data of same individuals in both the data sheets which reduced existing null values.
- As null values were further reduced and contributed very less percentage, imputing those null values was considered.
- The research paper by **Xueying Xu et al. (2020)** was used for the null value imputation techiques under which mice imputation led to the least bias. Hence, techniques such as Multiple Imputations (MICE) and imputation by KNN were adopted.
- 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 sckit-learn packages (label encoding).
- Apart from above methods a different approach was explored in null value imputation, where modelling was performed on the data.
- Encoded data was sorted into two not-null values and null values for a particular feature whose null value were to be imputed.
- Gradient boost algorithm was adopted in imputing where mean of all training f1 scores for all the features models were found as:
    - 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:

- Barplots
- Count plots with hue
- 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)

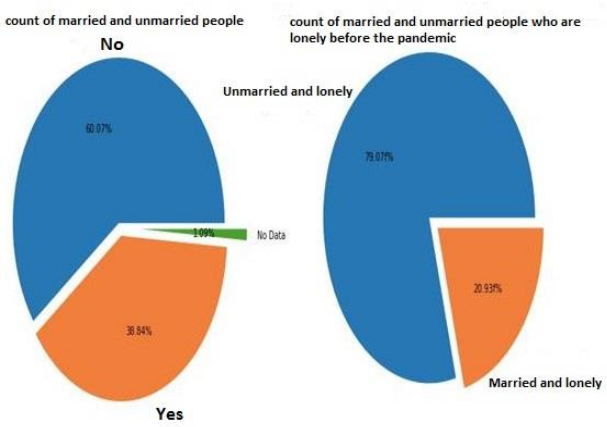


Fig (5.1.1)

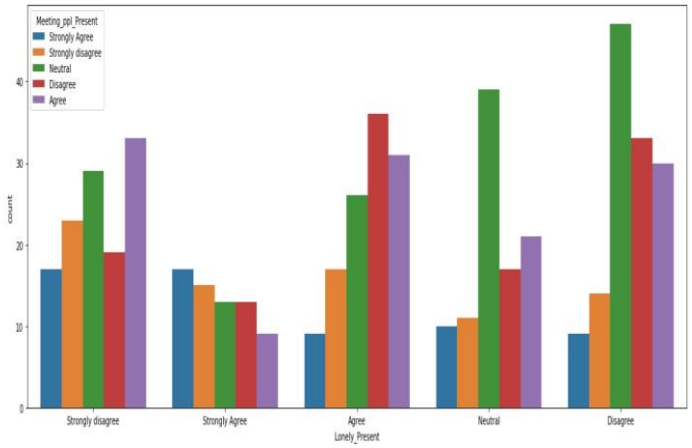


Fig (5.1.2)

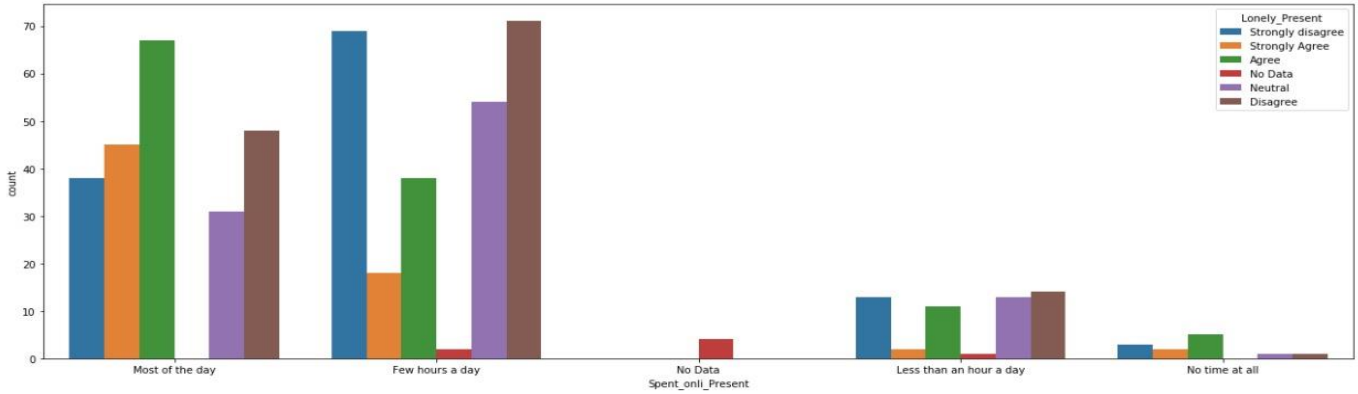


Fig (5.1.3)

Fig(5.1.1), Fig(5.1.2) and Fig(5.1.3) show an example of behavioural analysis and explains below inferences:

- 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.

5.2.Multi-Collinearity:

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 qualitative variables. The correlation between the independent variables has therefore been visualised using this technique in R.

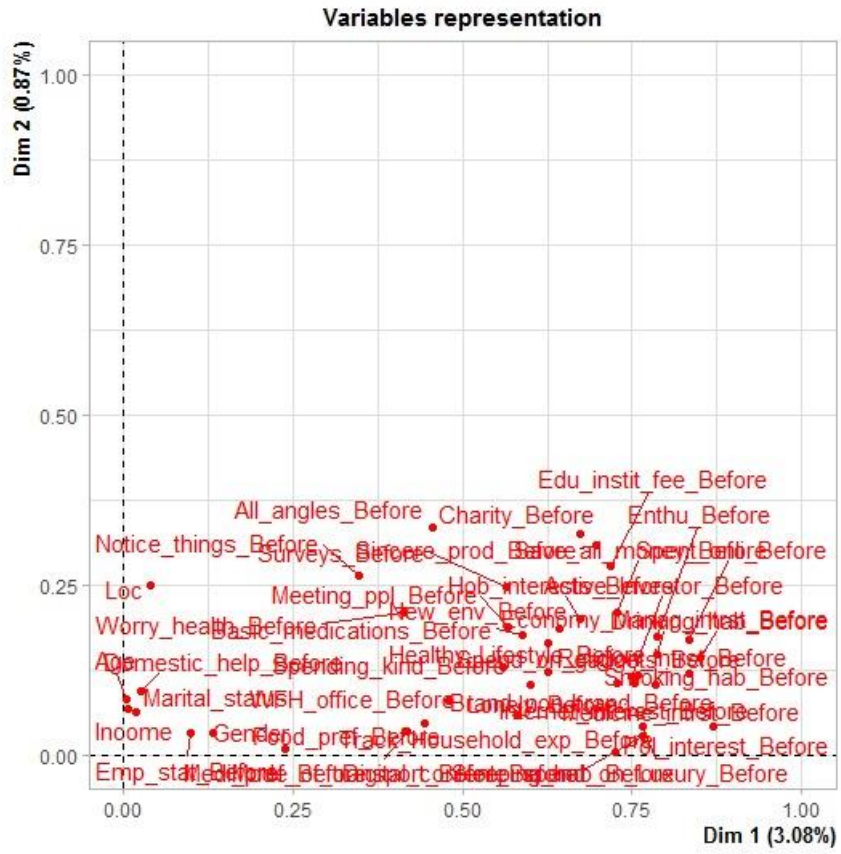


Fig (5.2.1)

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

5.3.Distribution of variables:

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

5.4.Presence of outliers and its treatment:

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.

5.5.Statistical significance of variables:

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 (9.5.1)

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 (9.5.2)

Fig (9.5.1) 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 (9.5.2) gives us the features which have significant change in their responses between before pandemic and current situation.

9.6.Class Imbalance and its treatment:

The dependent variable (spending kind) is a multi class variable with 3 classes (high, medium, low) and is highly imbalanced with the classes in the ratio, this will be treated using appropriate sampling techniques (SMOTE, NEAR MISS).

10.Feature Engineering:

10.1.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 sckit-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.

10.2.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 (10.2.1)

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 (10.2.2)

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

Fig (10.2.2) shows us the features which exhibit significant relationship with the target variable in current situation responses.

10.3.Dimensionality reduction:

Another issue to deal before modelling is the multi-collinearity and auto-correlation in the data. To deal with such issue MCA, a technique which reduces the multi-collinearity effect as well as reduces dimensions, was adopted.

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.

11.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.

11.1. 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 (11.1.1)

	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 (11.1.2)

Fig (11.1.1) and (11.1.2) 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%.

11.2. Other Models

11.2.1K Nearest Neighbor

kNN stands for k Nearest Neighbor. In data mining and predictive modeling, 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.

11.2.2 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  1]]
```

Fig(11.2.2.1)

11.3 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(11.3.1)

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.

12.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

12.1.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(12.1.1))
- Present dataset: Silhouette score of 0.26 with five clusters, with an inertia of 1055 were seen.

Fig(12.1.1)

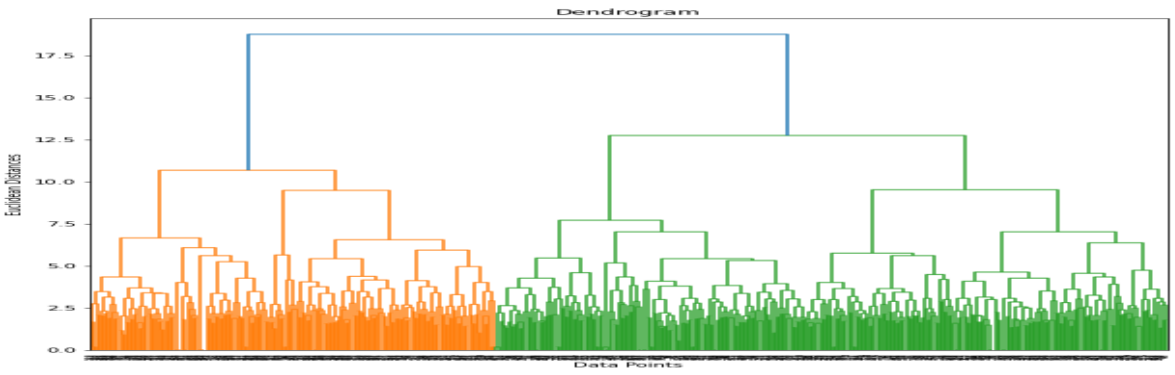
12.2 Hierarchical(Agglomerative) Clustering

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(12.2.1)

12.3 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 inorder to cluster categorical data.

The cost in K-Medoids algorithm is given as

$$c = \sum_{Ci} \sum_{Pi \in Ci} |Pi - Ci|$$

Fig(12.3.1)

The K-Medoids clustering when applied to the multivariate catgeroical 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.

12.4 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.

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(12.4.1)

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(12.4.1))
In the before and present situation(Refer Fig(12.4.1)

- 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)

13. 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

14.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

15.APPLICATION

- Educational Institutions: Analyse the change in student's behaviour in order to improve their performances and accordingly suggest reforms in the education system.
- Private Sectors: 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)