**Data Analysis**

***I should write how these variables are proposed in Venkatesh shit affect the variable***

There are 6 moderated variables which are “Age”, “Gender”, “Marital Status”, “Education Level”, “Work Industry”, “Work Position”.

Age

|  |  |
| --- | --- |
|  |  |

The preliminary analysis done for the “Age” Variable indicates a fairly even distribution between the age range of “< 25 years”, “26 – 40 years”, “41 – 55 years” and “above 55 years”. This is primarily because the survey were targeted for university students and adults whom have working experience. The descriptive statistics indicates that 29% of respondents were below 25 years old, 33.22% were around 26 – 40 years and 30.07% were around 41- 55 years. Based on the percentage result, it is safe to assume that there won’t be any bias based on age group, as the percentage for each age group is roughly the same and is evenly distributed.

When comparing age groups with the targeted variable of “In the next six months, do you plan to purchase anything using the E-payment mode?”, the graph indicates that majority of each respondents in each age group are highly likely to adopt e-payment.

Gender

|  |  |
| --- | --- |
|  |  |

The preliminary analysis done for “Gender” Variable indicates that among the 286 respondents, a total of 59.4% are male, whereas 40.6% are female. The respondents are fairly distributed among both genders, albeit a bit bias towards male. Therefore, before beginning our analysis, we may need to balance out between both genders.

When comparing genders with the targeted variable of “In the next six months, do you plan to purchase anything using the E-payment mode?”, the graph indicates that majority of each respondents in each age group are highly likely to adopt e-payment.

Marital Status

|  |  |
| --- | --- |
|  |  |

The preliminary analysis for “Marital Status” indicates that there is a fair distribution of respondents being either Single or Married. Among 50% of the respondents are single, whereas 47.55% of the respondents are married.

Education Level

|  |  |
| --- | --- |
|  |  |

There are a total of 4 different categories for Education Level, that is, “Primary School”, “Secondary / High School”, “College / University” and “Graduate School”. From here, we can see an uneven distribution among the respondents. As the questionnaire was mainly distributed in UTAR Campus, we can observe that majority of the respondents are categorized at “College / University” or ‘Graduate Level” at 63.38% and 16.08% respectively.

Work Industry

|  |  |
| --- | --- |
|  |  |

From an initial Analysis

Work Position

|  |  |
| --- | --- |
|  |  |

From a preliminary analysis, there are a total of 4 different categories of Work Position, that is “Junior Management”, “Middle Management”, “Top Management” and “Professional”. Among the respondents, 24.48% are employees in “Middle Management”, 16.43% are employees in “Professional” Position, whereas 11.19% are in “Junior management” position. *Effects of Work Position by Venkatesh*

**Correlation Among Features**

Correlation analysis is a technique of analyzing the linear relationship between two variables. The two variables can be independent or dependent based on the strength of the relationship computed. We define the strength of the correlation as correlation coefficient. Correlation coefficient may be derived from various formula such as Pearson’s Correlation coefficient, Spearman’s Correlation coefficient and many more.

The importance of correlation analysis is to determine the strength of relationship between two variables. Normally, it provides inside when trying to analyze causality in a dataset. Based on the strength and value of the correlation coefficient, we are able to determine whether the variable positively or negatively affect the predictor attribute.

One of a Hypothesis we commonly have is that attributes that rank higher will have higher effect in classification accuracy. Therefore, correlation analysis shall be conducted to identify the important attributes that increases the classification accuracy and filter out irrelevant features. The features that are selected may contain an underlying factor, that we shall investigate in our project as well.

To undergo such experimentation, we use two correlation coefficient methods, that is Pearson’s Correlation and Spearman’s Correlation Coefficient.

**Why do we use Correlation?**

The main issue is that if we use another metric like information gain, we would not be able to select features that have high correlation. Assuming our initial hypothesis is correct, it would mean that highly correlated features are the features that have the underlying meaning.

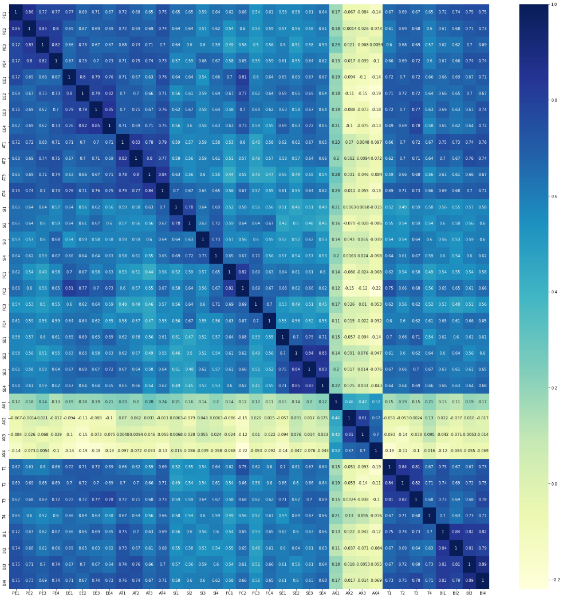
Diagram

Description automatically generated

**Pearson’s Correlation Coefficient**

***I Should put formula here, and talk about how to interpret the result of Pearson’s Correlation***

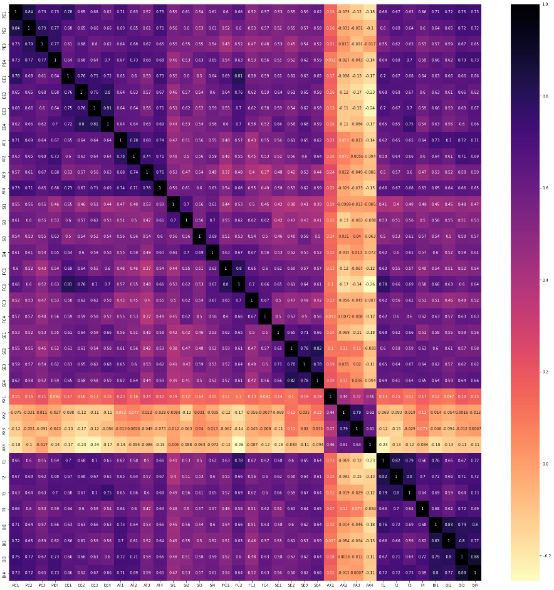
The table of computed result of UTAUT Factor Variables using Pearson’s Correlation:



**Spearman’s Correlation Coefficient**

***I Should put formula here, and talk about how to interpret the result of Spearman’s Correlation***

The table of computed result of UTAUT Factor Variables using Spearman’s Correlation:



Using both metrics of Pearson’s Correlation and Spearman’s Correlation, we ran a pairwise correlation among all 32 UTAUT Factors with the intention of discovering underlying relationship among the UTAUT Factors.

Coincidentally, the diagram of the heatmap of Pearson’s Correlation and Spearman’s Correlation are identical.

Factors AX1-4 are uncorrelated and not significant with other variables with the exception of itself.

The color shadings are much denser between PE, EE and AT indicating that these groups of factors shares an interesting relationship.

Although, we have made some interesting observation of the relationship between the UTAUT Factors, we still cannot infer any significant relationship. This is because a heatmap can only tells us about the strength of the relationship between two variables. It cannot show us a larger picture of how interconnected the factors are. Therefore, in this project, we also explore how to visualize correlation using network graphs.

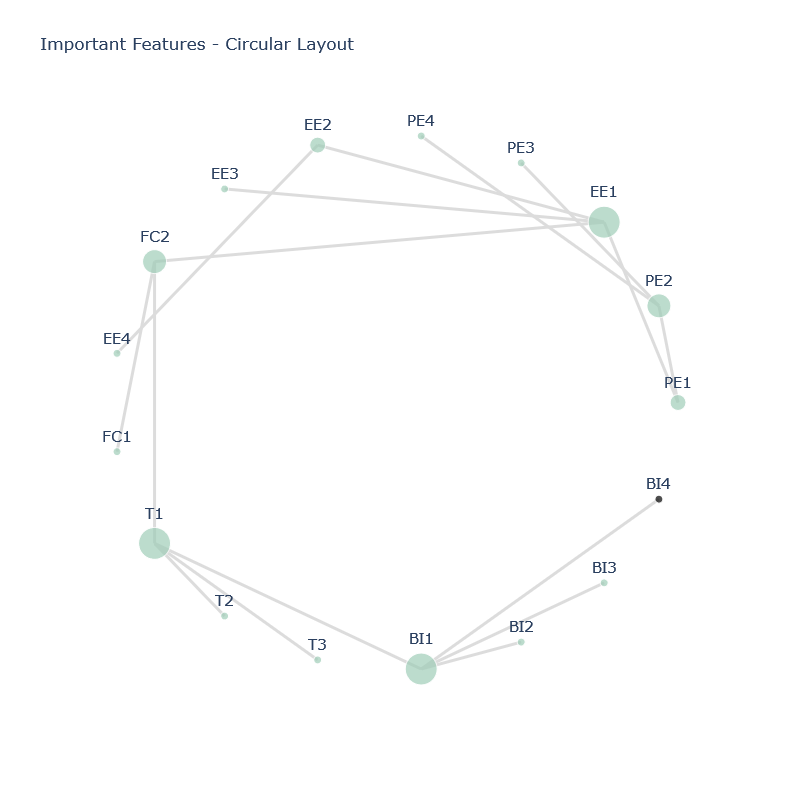
**Correlation Network Graph**

Network graphs is a mathematical structure to show relations between points in a less statistical manner. It allows us to study relationship between factors in a more aesthetically pleasing manner.

Network graphs are made up of nodes and edges. Commonly, Nodes or vertices are the discrete entities of the graph or dataset. Edges or links are used to represent relations among the nodes. Weighted edges can be used to represent correlation coefficient. In this scenario, a node represent a feature in the dataset, which is a UTAUT Factor, whereas Edges are used to represent the strength of correlation between the two UTAUT Factors. A positive relationship between the two factors is colored green whereas a negative relationship is colored red. The strength of the correlation is represented by the size of the node. A node that is larger in size also represents that it has a large degree, thus it affects numerous factors in the dataset.

In network analysis, there are two types of graph used to visualize correlation matrix. Circular layout and Fruchterman-Reingold Layout.

**Circular Layout**

****

Compared to heatmap, the circular layout instantly shows us which features are heavily interconnected. The features are less clustered and they are not defined by a statistical measure. To analyze the circular layout graph, features that are closed in proximity means that they are significantly correlated. The size of the Node represents the number of nodes it is connected to. The larger the node, the more connected it is to other features. The issue with the circular layout is that, due to the features having to form a circle, it does not create clusters of features. The size of the node is covered by the overall shape of the graph.

**Fruchterman-Reingold Layout**

Chart, radar chart

Description automatically generated

The fruchterman-reingold-layout removes the unnecessary circle layout from the above graph and depict it as a network graph. By using a minimum spanning tree structure, we are able to remove unnecessary edges from the graph and simplify it. The fruchterman-reingold layout also groups features into clusters. From here, the size of the node is more apparent, and represents how many nodes it is connected to. We gain more information from the fruchterman-reingold layout rather than the circular layout.

One of the more common things to measure in a network is centrality. Here we can observe that, EE1 is the central node, as the size of the node is the largest. It acts as a gateway between PE and FC Factors. T1 is also another central node that connects between FC2 and BI1.

From an empirical viewpoint, PE1, EE1, FC2 and T1 are the more prominent features as they are connected to more features. These factors may have an underlying factor as they are highly correlated.

**Feature Selection Methods**

Feature selection is a technique categorized as data pre-processing, that is a necessary step in determine important features. The main aim of feature selection is to determine the minimum number of feature subsets from a problem domain while retaining a high classification measures in representing the original features.

When the number of feature selected is rather small, the chances of information content may be low. On the other hand, the presence of noise as irrelevant data will be highly proable when many features are selected. Therefore, feature selection should be on the right selection of subsets, avoiding too large or too small number of features.

The performances of feature selection method is usually evaluated by the machine learning algorithm. A good feature selection method should have high learning accuracy but less computational overhead. Feature selection is a technique that has the ability to decrease the number of attributes by eliminating the least significant features.

The aim of this study is to find highly sets of correlational user behavior

There are four basic steps in a typical feature selection process. The process of feature selection is as below:

* The generation procedure is to generate the next candidate subset from the original feature set
* The evaluation function to evaluate the subset to determine the relevancy towards the classification task using measure for instances distance, dependency, information and consistency.
* Stopping criteria to decide when to stop. This is where it determines the relevant subset or optimal feature subset
* Validation procedure is to check whether the selected feature subset is valid.

**Filter Methods Data**

* **Info Gain**
* **Correlation Feature Selection**

**Wrapper Methods**

* **Forward-Selection**
* **Backward-Selection**

**Potential Target Variables**