**MS5105 Statistical Techniques for Business Analytics 2022-2023**

**Group 24 Assignment**

In submitting this work, I confirm that it is entirely my own. I acknowledge that I may be invited to an online interview if there is any concern in relation to the integrity of my submission, and I am aware that any breach will be subject to the University's Procedures for dealing with breaches of Exam Regulations. I am aware of what the NUI Galway plagiarism policy entails.

**Problem 1** (rottenapple.sav) RottenApple Inc. is experimenting which of two background colours for computer text is easier to read, as determined by the speed with which a task described by the text is performed. The study randomly assigned 35 employees to one of two versions of a computer program that presented text describing which of several icons the user should click on. The program measured how long it took until the correct icon is clicked. This measurement is called “reaction time” and is measured in milliseconds (ms). The two versions of the program differed in the background color for the text (yellow or cyan).

**Answer:** The dataset contains three columns ID, color, and reaction time respectively. So before jumping into the solution we will do a preliminary analysis. Let’s look at some of the descriptive and normality of the categorical and continuous variables which are color and reaction time respectively.

## **Descriptive Categorical variable (Color) gives quick look at the data:**

Table

Description automatically generatedTable

Description automatically generated***Fig 1.1 Color Statistics Fig 1.2 Color Distribution***

From Fig 1.1 we can see we do not have any missing values, and all are valid.

From Fig 1.2 we are able to see the distribution of color among the employee and we can see 17 which is 48.6 percent assigned to yellow and the remaining 18 which is 51.4 assigned to cyan which shows the distribution has taken almost equally.

## **Validating the Normality of the Continuous variable reaction time**

Table

Description automatically generated

The Kolmogorov-Smirnov shows significance = 0.2 which is greater than 0.05 and shows there is no significant difference in the reaction time for choosing a color.

* Taking a look at the histogram we can see it’s bell-shaped with no dips in the center.
* The Q-Q plot of the reaction time shows there is no significant difference between the Expected value and the observed value. Even though there are a few deviations but is not significantly far from the center line which implies normal.
* The box plot shows a low value of 300 and a high value of 1000 without any IDs present so no further analysis is needed

***Fig 1.3 Histogram Fig 1.4 Q-Q plot***

Chart, bar chart, histogram

Description automatically generatedChart, line chart

Description automatically generated

Table

Description automatically generated ***Fig 1.5 Meana and 5% Trimmed Mean Fig 1.6 Box Plot***

Chart, box and whisker chart

Description automatically generated

* From the descriptive of the continuous variable, we can see the reaction time has a Mean = 670.03 and 5% trimmed = 670.88 which is fairly close
* It has a Skewness = -0.116 which means its skewed towards the right and Kurtosis = -.848 which implies it’s a flat curve

## **Measuring the reaction time of the employees on choosing either yellow or cyan**

We will be using the T-test as our statistical tool in order to analyze a categorical variable Color Vs a continuous variable Reaction time. (Analyze > Compare Means > Independent Samples T-Test)

***Fig 1.6 Levene’s significance***

Table

Description automatically generated

Levene’s significance = 0.272 which is greater than 0.05 which means there is no significant difference.

From the group statistics, we can see that,

Table

Description automatically generatedFor Yellow 🡪 Mean = 679.65 and SD = 159.38

For Cyan 🡪 Mean = 660.94 and SD = 202.03

Even though there are some differences, there is no significant difference in the mean and SD for yellow and cyan.

The t-test shows t(33) = 0.303 and t(32.01) = 0.305 which has a significant value of 0.764 and 0.762 respectively which is greater than 0.05 which means there is no significant difference in the employees choosing a text based on its color.

**Final thoughts:** Based on our analysis we can see there are no missing values based on the descriptives of a categorical variable, the continuous variable reaction time is normally distributed and the reaction time in choosing a text-based did not significantly differ based on the text color.

**Problem 2** (sales.sav) In a small retail business there are three clothing stores. Each store sell three types of products: budget items, mid-range, and high fashion. The business manager has decided to investigate if sales of these types differ by store. To do this, the manager retrieved a list of sales at all three stores and created a dataset also recording which store sold which item.

## **Descriptives of categorical variables**

Answer Before we start the analysis, we will look at the descriptives which give a piece of quick information about the data. We will not be checking for normality as both variables are categorical and not continuous.

Table

Description automatically generated

Based on the statistics, we can see there are no missing values, and all are valid. And based on the frequency table given below for store and product type, we can infer that the distribution has not taken place evenly. So, we will analyze this data to check how the sale is differing based on the store location.

Table

Description automatically generatedTable

Description automatically generated

## **Analyzing if the product sold differ based on the store**

The given sales dataset contains the sales of three different items across three different stores. Since we are considering two categorical variables store and product, we are utilizing the chi-square test to figure out which store is selling which item and if there is any difference in sales based on the store. (*Analyze > Descriptive Stats > Crosstabs*)

**Fig 2.1 Summary data of Store Vs Product**

Table

Description automatically generated

From Fig 2.1 we can see there are no unsolved cases, so all 270 cases have been addressed. The cross tabs show the proportion of different products like budget items, mid-range items, and high fashion across 3 different stores.

From Fig 2.2 we can see the Pearson Chi-Square value = 0.001 which is less than 0.05 which signifies there are significant differences. We do not need Yate’s Continuity as it’s not a 2x2 table. Since p (0.001) < 0.05 means the products sold are significantly different based on the store. The Phi and Cramer’s V test also shows p=0.001 which is less than 0.05 which supports the above theory.

**Fig 2.2 Pearson Chi-Square Fig 2.3 Phi Cramer’s V Values**

Table

Description automatically generatedTable

Description automatically generated

From the store Vs Product type cross-tabulation, we can infer the below,

1. The East Outlet Stores has sold Budget =26 items, High Fashion = 48 and Mid-Range = 30 which come to a total of 104 products.
2. The South Outlet Stores has sold Budget = 26, High Fashion = 46 and Mid-Range = 44 which come to a total of 116 products
3. The West Outlet Stores have sold Budget =26, High Fashion = 12 and Mid-Range = 12 which comes to a total of 50 products

**Fig 2.4 Products sold across three different stores.**

Table

Description automatically generated

So based on the cross-tabulation, The west outlet stores have sold the least number of items for all the three different types of products. Even though the east outlet and south outlet stores sold the same number of items in Budget product, the east outlet sold more items in high-fashion and mid-range compared to both south-outlet and west outlet with the highest total of sol products = 116 which implies there is a difference.

**Final thoughts:** So, based on the tests performed Pearson chi-square,Phi and Cramer’s V show there is a significant difference and to furthermore support the theory, the cross-tabulation clearly shows there is a significant difference in the products sold based on the store location

**Problem 3** (holidays.sav) A company wants to investigate the differences between number of annual holidays taken by its employees of three departments. For this purpose, an HR manager takes random sample of 40 employees from each department. The manager creates a dataset by extracting, from HR records, the number of annual holidays during last year for each employee in the sample.

**Problem 4** (starcraft.sav) Starcraft is one the most popular real-time strategy game. During one of its tournament, the results of games played by 100 randomly selected players were recorded. The dataset contains the army and strategy used by each player along with the percentage of wins. We are interested in finding out which combinations of army and strategy are the most effective.

Answer The given Starcraft dataset contains three columns wins, army, and strategy which are continuous, categorical, and categorical with more than 2 levels respectively. The dataset contains the number of wins by a player in 100 games using which type of army and strategy. Since this contains 2 categorical variables with more than 2 levels and 1 continuous variable, we will use 2-Way Anova to figure out which combinations of army and strategy were effective.

## **Analyzing the dataset using 2-Way Anova**

Table

Description automatically generated***Fig 4.1 Levene’s Test for Significance***

***Fig 4.2 Interaction based on Significance***

Table

Description automatically generated

From Fig 4.1 we can see Levene’s test has significance = 0.156 which is greater than 0.05 which means there is no significant difference in the values and the test of between-subjects effects show the interaction if any.

1. The significance of Army F (2,100) = 0.694 and p=0.502 which is greater than 0.05 which is insignificant
2. The significance of Strategy F (3,100) = 0.852 and p=469 which is greater than 0.05 which is insignificant
3. The significance of Army\*Strategy F (6,100) = 0.462 and p=0.835 which is greater than 0.05 which is insignificant

So, the number of wins by the player doesn’t vary much based on his army or strategy or the combination of both. From Tukey’s Post Hoc test we can see the strategies Balanced, Build, Defend, and Rush doesn’t significantly differ from each other.

***Fig 4.3 Tukey’s Post Hoc test***

Table

Description automatically generated

**Final thoughts:** From Levene’s Test and Post Hoc Test we can see there is no significant difference. Hence, we were not able to see any difference in the win percentage by the player based on the selection of Strategy and Army

**Problem 5**(crimes.sav) Crime statistic were recorded for a random set of communities in a large city. The data includes the number of violent and non-violent crimes in a month. The city administrators are interested in understanding whether number of residents and average value of houses in a community influence both types of crimes; and what is the relationship between violent and non-violent crimes.

Answer

We will be utilizing the following techniques and technologies in order to comprehend and provide a good representation of the data for the understanding of the target audience.

1. XL Data Mining – To eliminate redundant data, NULL values, identify outliers, and create dummies to address redundancy, we must preprocess the data.
2. SPSS (Statistics Tool) – We will utilize the Statistics tool in order to check the data normalization if there is any correlation between the data and to point out any significant differences if any
3. Tableau – The preprocessed data will be used by Tableau for exploration, trend detection, and insight sharing, providing a detailed analysis that will be displayed through an interactive dashboard and visualization.