|  |
| --- |
|  |
| Urban Renewal Planning  Technical Report |
| |  |  |  | | --- | --- | --- | | Soe Htet | 9/18/22 | Data Analytics Lab | |

**Contents**

[1. Overview 2](#_Toc114178199)

[2. Objective 2](#_Toc114178200)

[3. Data Preparation 2](#_Toc114178201)

[3.1 Wage data is recorded with significant outliers. 2](#_Toc114178202)

[3.2 Duplicate rows 3](#_Toc114178203)

[3.3 Timestamp column is in wrong data format 3](#_Toc114178204)

[3.4 Amount column is transformed into a more accessible format for analysis 3](#_Toc114178205)

[3.5 Participant table joined with financial data tables for detailed analysis 4](#_Toc114178206)

[4. Data Analysis 4](#_Toc114178207)

[4.1 Trend of recreational spending 4](#_Toc114178208)

[4.2 Wage Vs Education 8](#_Toc114178209)

[4.3 Correlation between food spending, shelter spending & recreational spending 10](#_Toc114178210)

[4.4 Financial health of participants during survey period 12](#_Toc114178211)

[4.5 People of different ages have different interest groups. 13](#_Toc114178212)

[4.6 Joviality is inversely proportional to income level. 15](#_Toc114178213)

[4.7 Having kids contribute to happiness of the family. 17](#_Toc114178214)

[5. Interpretation of Analysis Results 18](#_Toc114178215)

[6. Recommendation for Further Analysis 18](#_Toc114178216)

# **1. Overview**

Over the last decade, technology has become an essential part of people’s daily life and enormous information of digital footprint has been generated regarding spending patterns, income, household demographics and many other things. Consequently, town planners have been given valuable access to wealth of data which can be used to gain crucial insights to assist in community redevelopment efforts.

# **2. Objective**

The objective of this study is to analyse the particulars and financial survey data of 1000 participants, present insights on resident’s wage, spending trends, overall financial health and joviality of the population and use these insights to assist city revitalization plan and allocate city renewal funds.

# **3. Data Preparation**

Two survey datasets named ‘Particulars.csv’ and ‘FinancialJournal.csv’ were used in this analysis. Participants dataset includes participant’s ID, education, household size and joviality data while Financial Journal dataset includes daily financial records of education spending, food spending, shelter spending, rent adjustment, recreation spending and wage.

Data preparation and analysis of survey data were performed on statistical analysis software JMP Pro 16.

The following sections explain the steps taken for data cleaning, transformation and combining to create clean datasets for subsequent analysis.

## **3.1 Wage data is recorded with significant outliers.**

Wage data in financial journal survey has significantly high numbers of large values in box plot graph. Upon further analysis, wage data at the start of the survey (1st day of March) is found to be recorded in thousands ($) as opposed to daily wage data that should be in hundreds typically, similarly to the rest of the survey data. The outlier data were selected and excluded from subsequent analysis.

Figure 1: Box plot of amount ($) recorded by category

Chart, box and whisker chart

Description automatically generated

## **3.2 Duplicate rows**

1113 numbers of rows in financial data survey are duplicates. They were also excluded from the analysis to maintain the integrity of the dataset.

## **3.3 Timestamp column is in wrong data format**

Date and time values in timestamp column were recorded as character data type. This will hinder the ability to analyse timestamp data and therefore, a new column timestamp recoded was recreated to store values in y/m/d h:m format.

## **3.4 Amount column is transformed into a more accessible format for analysis**

Amounts (spendings & wage) are recorded daily in financial journal. It is not convenient for analysis as the amounts entered in a row can be related to different categories (wage, shelter, food, etc). Thus, tabulate function was used to format the data to cleaner and more understandable layout for further analysis. Two new data tables were created as shown below. The first table documents overall spending and wage received across the duration of the survey of each participant while the second table shows the data table that is further categorized by months.

Table 1: Sum of amounts by category

Table

Description automatically generated

Table 2: Sum of amount by category by month

Table

Description automatically generated

## **3.5 Participant table joined with financial data tables for detailed analysis**

Two new data tables named financial data (overall) and financial data (monthly) were created by joining data tables. They will be used for in depth data analysis and hypothesis testing.

# **4. Data Analysis**

After cleaning and transforming the survey data sets in section 3, the results and trends gained from exploratory data analysis are discussed in this session.

## **4.1 Trend of recreational spending**

**Data insight 1** - Recreational spending has clear downward trend from month 3 to month 8.

To understand and explore the spending patterns of participants in the survey, firstly, recreational spending is plotted in a line chart with months as independent X variable. As shown in figure 2, it is clear that mean values of spending on recreation are steadily declining and exhibiting a downward trend from month 3 to month 8.

Figure 2: Trend of recreational spending from month 3 to month 8

Chart, line chart

Description automatically generated

Table 3: Summary statistics table of recreational spending by month

Graphical user interface, text, application, table

Description automatically generated

To choose the hypothesis testing method in confirmatory data analysis, normality of recreational spending data is tested by using the following hypothesis.

Null Hypothesis H0: Recreational spending distribution is sampled from normal distribution.

Alternative Hypothesis H1: Recreational spending distribution is not sampled from normal distribution.

Confidence interval is set at 95%.

Figure 3: Normality assumption test on recreational spending by months

A screenshot of a computer

Description automatically generated with medium confidence

As seen from goodness of fit tests of the figure above, both Shapiro-Wilk test and Anderson-Darling test return p-value less than 0.05. Therefore, there is significant statistical evidence to reject the null hypothesis that these distributions are sampled from normal distribution. Thus, non-parametric tests using Wilcoxon Method will be employed for means comparison.

Null hypothesis and alternative hypothesis are as follows.

Null Hypothesis H0: The difference of the means of the distributions is equal to zero.

Alternative Hypothesis H1: The difference of the means of the distributions is not equal to zero.

Figure 4: Nonparametric comparison of means of recreational spending

Chart, box and whisker chart

Description automatically generated

From the figure above, p-values from each pair comparison are all below significance level of 0.05 except for last 4 pair comparisons. Thus, there is statistical evidence that means of recreational spending of months 3 and 4 are statistically different from the rest of the months. However, for month 6,7 and 8, the results failed to reject null hypothesis, therefore, it cannot be concluded that there is significant difference among the means of recreational spending of those months.

## **4.2 Wage Vs Education**

**Data insight 2** - Higher education level commands higher wage.

From the box plot analysis of wage data categorized by education level, it can be seen that there seems to be a correlation between participants getting higher wage with higher education level.

Figure 5: Monthly Wage vs Education Level

Chart, box and whisker chart

Description automatically generated

It can be inferred from the graph that participants with graduate degree are drawing highest median wage of $5375 followed by bachelor’s degree holders with median wage of $4291.

Normality test is performed for monthly wage data with confidence interval of 95%. P-value less than 0.05 from Anderson-Darling test concludes that this observed distribution does not conform to a normal distribution.

Figure 6: Monthly Wage Distribution

Graphical user interface, application

Description automatically generated

To confirm with statistical methods that mean values of wage data by education are indeed different, nonparametric test is carried out with significance interval of 0.05.

Null hypothesis H0: Means of wage data for all education levels are the same.

Alternative hypothesis H1: Means of wage data for all education levels are not the same.

Figure 7: Nonparametric comparisons of means of wage by education level

Box and whisker chart

Description automatically generated

P-values of all pair comparisons fall below the significance level of 0.05. Therefore, null hypothesis that implies means of wage data for all education levels are the same is rejected as there is significant statistical evidence against the null hypothesis.

## **4.3 Correlation between food spending, shelter spending & recreational spending**

**Data insight 3** - Food spending is positively correlated to recreational spending

**Data insight 4** - Recreational spending is positively correlated to shelter spending

**Data insight 5** - Shelter spending is positively correlated to food spending

To investigate the possibility that participants who spend large amounts of money on recreation might also spend big on other categories of expenditures, scatter matrix plot is performed. As evidenced by the figure below, there is a moderately strong positive correlation between shelter spending, food spending and recreational spending.

Figure 8: Scatterplot matrix of shelter spending, food spending and recreational spending

Chart, scatter chart

Description automatically generated

Normality test is performed on all of 3 spending categories. Null hypothesis is rejected as the there is enough statistical evidence against it with p-values less than 0.05.

Figure 9: Normality test on food spending, recreation spending & shelter spending

Graphical user interface

Description automatically generated

To test the significance of correlation between these variables, hypotheses tests must be performed with appropriate significance level. In this test, alpha value is set at 0.05.

Null Hypothesis H0: The population correlation coefficient is not significantly  
different from zero. There is not a significant linear relationship between the two variables  
Alternate Hypothesis H1: The population correlation coefficient is significantly  
different from zero. There is a significant linear relationship between the two variables

Table 4: Hypothesis testing of correlations between spending variables

Table

Description automatically generated

It can be noted from the figure above that in both pairwise correlations test and nonparametric tests such as Spearman’s p and Hoeffding’s D, significant probability is found to be less than alpha value of 0.05 giving statistical evidence to reject the null hypothesis and conclude that there is a significant linear relationship between these variables.

## **4.4 Financial health of participants during survey period**

**Data insight 6** - Thriving economy is essential for the community’s well being and livelihood. It is important for the city planner to gain insights to residents’ financial well beings in order to effectively plan city development schemes. For this reason, exploratory data analysis is conducted on monthly saving trends of survey participants.

Figure 10. Mean and median of monthly savings by months

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

From the line charts above, there is a clear uptrend of monthly savings of the residents across the survey period. There is a slight dip in both mean and median savings on month 7. However, overall, it can be deduced that monthly savings have significantly increased over this period.

## **4.5 People of different ages have different interest groups.**

**Data insight 7** - Older participants have particular interests in interest group F if they don’t have kids and interest group D if they have kids.

**Data insight 8** - For younger participants without kids, there is no standout interest group. However, those with kids seem to hold special interest in interest group J.

Figure 12: Demographics of interest groups

Chart, waterfall chart

Description automatically generated

Table 5: Age vs Interest Group (Without kids)

Text, table

Description automatically generated

Table 6: Age vs Interest Group (With kids)

Text, table

Description automatically generated

## **4.6 Joviality is inversely proportional to income level.**

**Data insight 9** - Participants from lower income group reports higher satisfaction and happiness compared to those from high income group.

Although it is widely accepted that having money can solve most of the problems in one’s life, it does not always translate into being happy and content in life. Monthly wage data is divided into three income groups of low, middle and high income groups that have lowest 25%, middle 50% and highest 25% of monthly wages respectively. Figure below demonstrates that in fact, low income group reports highest joviality score followed by middle income group and then, high income group.

Figure 13: Joviality plot by income group

Chart, box and whisker chart

Description automatically generated

The distribution plots of joviality by income group are shown in figure below. It is apparent from the plots that the observed distributions do not conform to normal distribution. Therefore, nonparametric hypothesis testing is used to compare means.

Figure 14: Distributions of joviality by income group

Graphical user interface

Description automatically generated with medium confidence

For Wilcoxon hypothesis tests, null hypothesis and alternative hypothesis are defined as follows.

Null hypothesis H0: The differences in means of the distributions are equal to zero.

Alternative hypothesis H1: The differences in means of the distributions are not equal to zero.

Alpha level or significance level is set at 0.05 for this test.

Figure 15: Hypothesis testing of joviality by income group

Graphical user interface

Description automatically generated

The test results show p-value less than significance level of 0.05 for all comparisons. Therefore, there is statistical evidence to reject null hypothesis and conclude that means of the distributions are indeed statistically different.

## **4.7 Having kids contribute to happiness of the family.**

**Data insight 10** - Having kids increases happiness level of participants from low income group and high income group.

As derived from the plot below, having kids generally increases joviality score in low income and high income groups while it seem to not make a difference in middle income group.

Figure 16: Joviality by income group and having kids

Chart, box and whisker chart

Description automatically generated

# **5. Interpretation of Analysis Results**

1. Recreational spending has dropped substantially during the survey period of month 3 to month 8.

2. Participants who are big spenders in recreational activities are found to spend big as well in food and shelter.

3. Participants with higher education are drawing higher median salary. The city council can consider using part to city renewal grant funds to encourage residents to upskill themselves or pursue higher education.

4. The residents of the city are in great financial shape. Monthly savings has grown significantly with median monthly saving at month 8 becoming nearly 50% higher compared to month 3 at the start of the survey.

5. People with highest income have reported lower joviality score, perhaps due to stress and pressure from hectic lifestyle. City council may use renewal funds for mental health and well-being workshops and promote work life balance.

# **6. Recommendation for Further Analysis**

1. Household size data in the survey is only up to 3-person household. More data can be collected to include bigger household sizes for more complete demographic analysis.

2. Employment status of survey participants should be collected to better assess the state of current job market in the city.

3. Appropriate cluster analysis techniques can be used in future studies to form clusters of the population and analyse similar patterns or trends among the clusters.

4. In future surveys, we should collect opinions of participants on which facilities or infrastructure are due for a major upgrade. This data would be useful to carry out further analysis to assist in infrastructure upgrading plans.