RIT

STAT-614 Applied Statistics Project Report

December, 2022

Title:

Analysis on Impact of Age on Customer Buying Habits using Two-Sample Inference Testing

By

John Melwin Richard – 374009170

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
1	**INTRODUCTION**	3
2	**ANALYSIS** 2.1 UNDERSTANDING THE DATA 2.2 ASSUMPTIONS 2.3 HYPOTHESIS 2..4 TEST SELECTION 2.5 POOLED T-TEST 2.6 POOLED T-TEST CALCULATIONS	4
3	**CONCLUSION**	6
4	**RECOMMENDATIONS** 4.1 ANSWERS TO QUESTIONS 4.2 ANALYSIS USING JMP PLOTS	7
	REFERENCES	8

CHAPTER 1

1.1 INTRODUCTION

Social media platforms and websites have given us the competence to gather billions of data that take the guesswork out of digital advertising. Availability of tons of data has made it feasible for marketers to reach out to an ideal target audience using analytics and drive business outcomes.

One's socio-demographic aspects like age, gender, salary, and many other factors crucially influence consumer buying behaviour. This project uses the application of two sample inference testing to analyze an open-source dataset that contains demographics about customers who clicked a particular advertisement online. The conclusion and recommendations from this statistical analysis would assist the business in competing strongly in the market and make profit.

Dataset info : The dataset is obtained from Kaggle, a subsidiary of Google LLC. The raw data contains the following demographics about customers who clicked a particular advertisement online

1. '**User ID**': unique identification for consumer
2. '**Age**': customer age in years
3. '**Estimated Salary**': Avg. Income of consumer
4. '**Gender**': Whether consumer was male or female
5. '**Purchased**': 0 or 1 indicating purchased the product

1.2 OBJECTIVES

Gaining a cognizance of how age differences influence purchase decisions is pivotal for any business that wants to be successful in the market. Age is one of the decisive factors to let a person decide the way he/she wants to purchase a product. For instance, younger folks devote a greater share of spend to specialty beauty retailers compared to the mature population. Upon scanning the data, we have columns which gives us information about the age of the customers and also whether the customer has purchased the product or not. So, we would like to analyze the data and answer the following questions:

1. **Did age impact purchase behaviour?**
2. **Whether the mean age of customers who purchased the product is greater than who have not purchased? –** *Analysis*
3. **Which particular category of age group benefits most from the product?**

Furthermore, the data will be visualized using plots in JMP to better analyze and understand the impact of customer age against buying behaviour. Conclusions will be drawn based on statistical results as well as graphical inference.

CHAPTER 2

2. ANALYSIS

The two-sample inference testing will be used to analyze whether the mean age of customers who purchased the product is greater than who have not purchased the product.

2.1 UNDERSTANDING THE DATA:

The 'Purchased' column gives us the information whether a particular person has purchased the product or not and their corresponding age is available in the 'Age' column. '***Age'*** and '***Purchased***' column values are used as inputs for the two-sample inference testing.

- Purchased column: **0** -> Purchased the product (*population-1*), **1** -> not purchased the product

	User ID	Gender	Age	EstimatedSalary	Purchased
1	15566689	Female	35	57000	0
2	15569641	Female	58	95000	1
3	15570769	Female	26	80000	0
4	15570932	Male	34	115000	0
5	15571059	Female	33	41000	0
6	15573452	Female	21	16000	0
7	15573926	Male	40	71000	1
8	15574305	Male	35	53000	0
9	15574372	Female	58	47000	1
10	15575002	Female	35	60000	0
11	15575247	Female	48	131000	1

Figure 1. Demographics data of people who clicked the advertisement

2.2 ASSUMPTIONS:

1. Let $X_{11}, X_{12}, \ldots, X_{1n1}$ be a random sample from population 1(**Purchased = 0**).

2. Let $X_{21}, X_{22}, \ldots, X_{2n1}$be a random sample from population 2 (**Purchased = 1**).

3.The two populations X_1 and X_2 are independent.

4.Both X_1 and $X_2$are normal

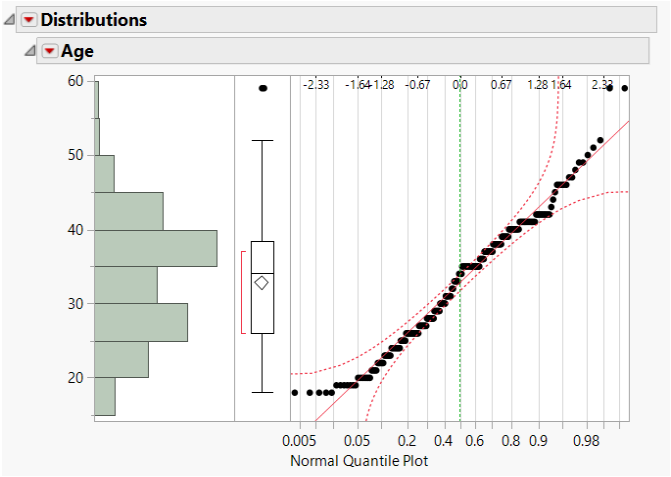


Figure 1. 1 Normal quantile plot for population 1(purchased =0)

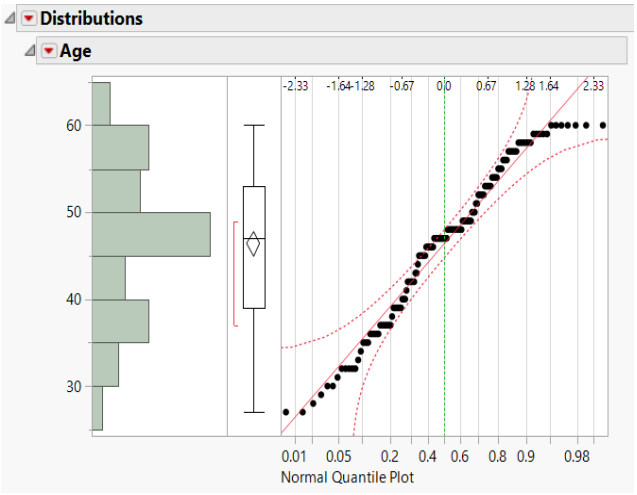


Figure 1. 1 Normal quantile plot for population 2(purchased =1)

- *From the above two normality plots the assumption 'Both X_1 and X_2 are normal' is verified.*

2.3 HYPOTHESIS

H_0(Null hypothesis): $\mu_1 = \mu_2$ *(mean age of people who purchased the product and who have not purchased the product are similar.)*

H_1(Alternate hypothesis): $\mu_2 > \mu_1$ *(mean age of people who purchased the product is greater than the mean age of people who have not purchased the product.)*

Level of Significance is $\alpha = 0.05$

2.4 SELECTING TYPE OF TWO SAMPLE INFERENCE

As the population variances are unknown here, we go for the t-test. The t-test for two means has two variants: (pooled t-test and Welch-Satterthwaite t-test). The sample variances are analysed below to decide between pooled t-test and Welch-Satterthwaite t-test.

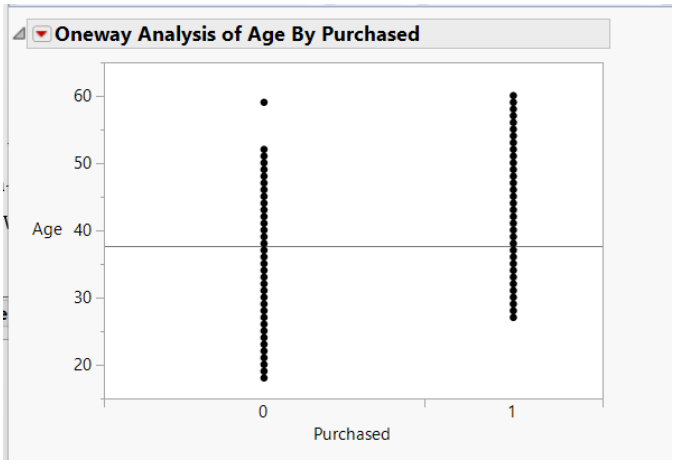


Figure 2. Analysis of variance of Age by Purchased

Assumption: From the above plot we could see the average age of customers who purchased the product is higher than who have not purchased however this could be verified statistically using the two-sample inference testing.

From the above plot, we can infer that there is not much significant difference in the variances of age between purchased and not purchased. Using the pooled t-test would be a suitable option because of similar variances.

2.5 POOLED T-TEST IN JMP:

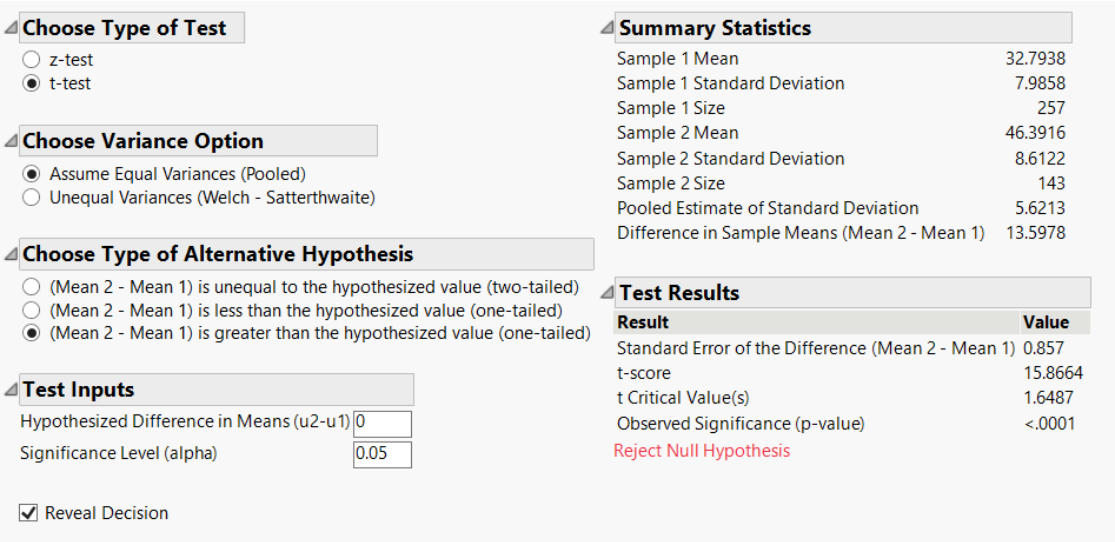


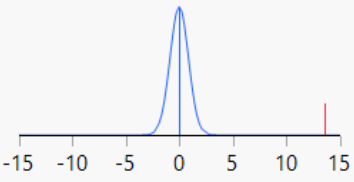
Figure 3. Pooled t-test (Age vs Purchased) on JMP

Pooled t Test

1-0

Assuming equal variances

Difference	13.5978	t Ratio	15.86638
Std Err Dif	0.8570	DF	398
Upper CL Dif	15.2827	Prob > |t|	<.0001*
Lower CL Dif	11.9130	Prob > t	<.0001*
Confidence	0.95	Prob < t	1.0000



2.6 POOLED T-TEST CALCULATIONS:

$\bar{x}1$ (Purchased = 1) = 32.793 $\bar{x}2$ (Purchased = 0) = 46.391

$S_1 = 7.985$, $n_1 = 257$ $S_2 = 8.612$, $n_2 = 143$

$$S_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$$

$$= \sqrt{\frac{(257-1)7.985^2+(143-1)8.612^2}{398}}$$

$$= 8.24$$

$$t_0 = \frac{46.391-32.793-(0)}{8.24\sqrt{\frac{1}{257}+\frac{1}{143}}}$$

$$= 13.597 / 0.857$$

$$= 15.8664$$

- Critical Value $t_{398,0.025} = 1.6487$, Reject Region is anything greater than 1.6487
- P-value < 0.0001

Test statistic t_0 = 15.8664

Rejection region: t_0>1.6487

p-value: <0.0001

α = 0.05

Decision: Reject the null hypothesis.

CHAPTER 3

3. CONCLUSION

Since the obtained p-value (<0.0001) is lesser than α (0.05), our decision is to Reject H_0 OR by rejection approach: the test statistic is more extreme than the critical value, so our decision is to Reject H_0. *We can conclude, at the 5% level of significance, that there is enough evidence to claim that the mean age of people who purchased the product is greater than the mean age of people who have not purchased the product. Also, the assumption from the age by purchased plot is now verified statistically*.

CHAPTER 4

4. RECCOMENDATIONS

As we have verified that the mean age of the people who purchased the product is greater than the mean age of people who have not purchased the product, we recommend to reach out more to mature people whose mean age is around 45 years. This specific group of audience are most likely to want to buy the product or service. Focusing on advertising target audience helps the business to develop an effective marketing strategy while saving time and money along the way.

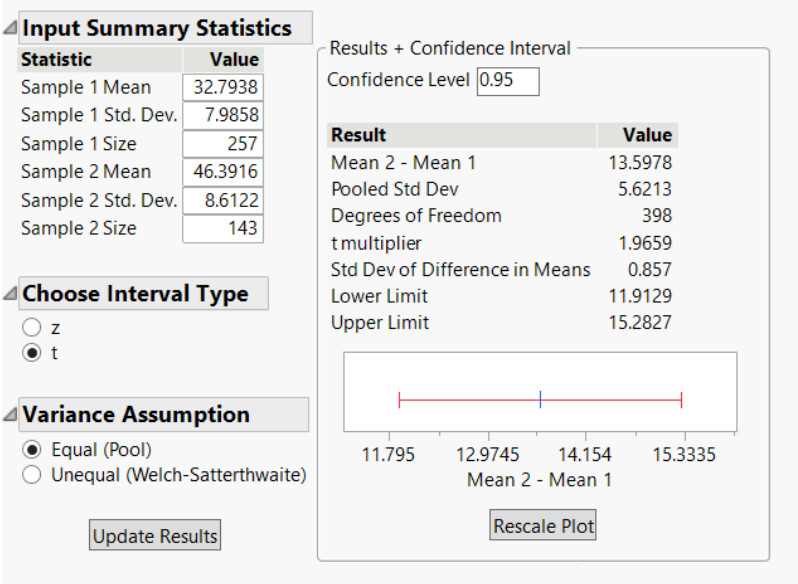


Figure 4. Confidence interval for mean difference on JMP

- 95% confidence interval for the difference of means: $11.912 < \mu_2 - \mu_1 < 15.2827$

Since the interval does not contain the value zero and positive, we would be able to verify that the mean age of the people who purchased the product is greater than the mean age who have not purchased the product.

There is a minimum difference of 11.912 and a maximum difference of 15.2827 between the means of population who have purchased the product and who have not purchased the product.

4.1 ANSWERS TO THE QUESTIONS:

1. **Does the age of the people impact buying behaviour?**
   Yes, the age of the people impacts buying behaviour maybe because of particular needs at their age.
2. **Whether the age of customers who purchased the product is greater than who have not?**
   Yes, it is verified statistically and graphically as well.
3. **Which particular category of age group benefits most from the product?**
   In this data, people between 45-50 age range benefits most from the product.

4.2 FURTHER ANALYSIS USING PLOTS AND RECOMMENDATIONS

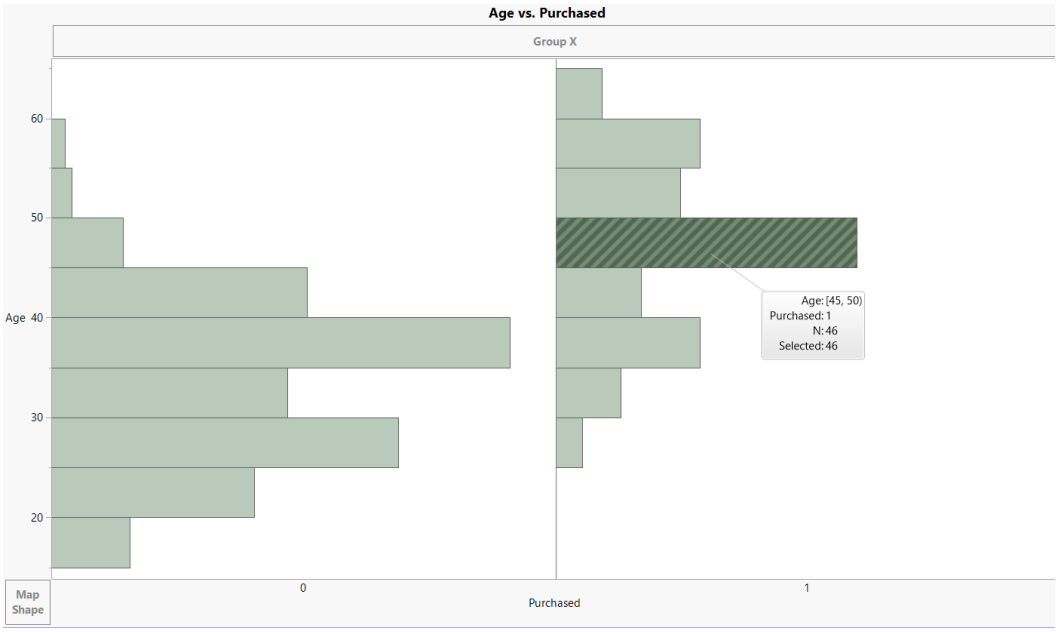


Figure 5. Histogram (Age vs Purchased) on JMP

- From the histogram above, we can infer that the people with age between 45 and 50 have maximum purchased the product. So, it is recommended to target people in this age range to advertise the product directly so that they will want to buy them.
- There is lot of people in the 35-40 age range who have not purchased the product. The company can work on covering this age range by modifying the product or releasing a different product line-up.

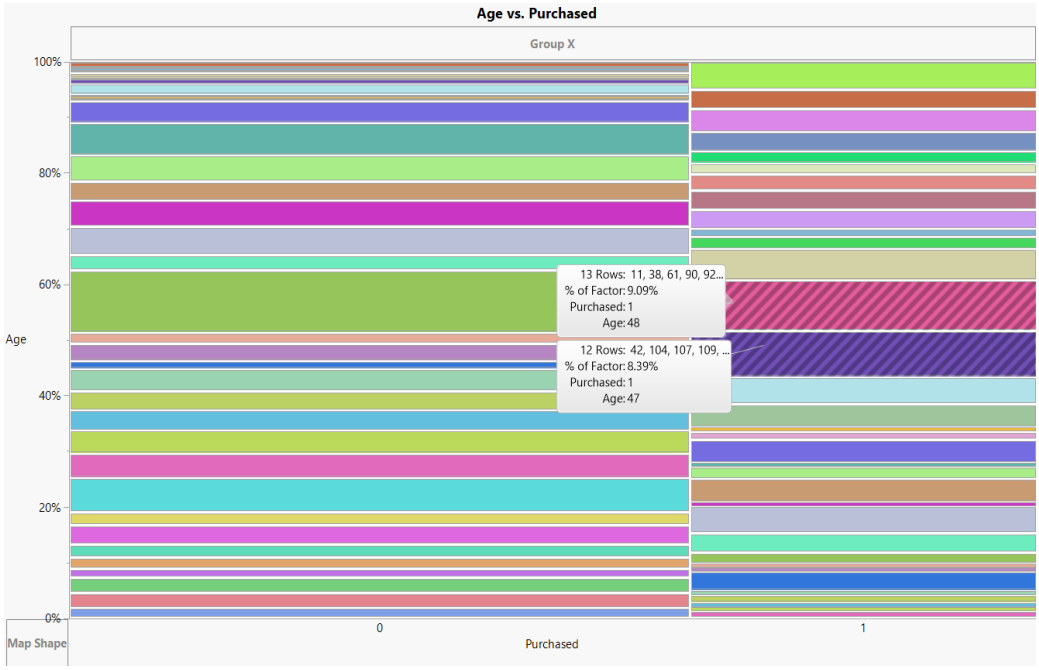


Figure 6. Mosaic plot (Age vs Purchased) on JMP

- From the Mosaic plot above, it is inferred that people with age = 48 and age = 47 have utmost brought the product. Using this information, the company can specifically utilize this age groups to understand whether they are recurring customers.

REFERENCE:

Dataset link: https://www.kaggle.com/datasets/rakeshrau/social-network-ads