A Case Study Report on

A/B Testing for Business Growth: Analyzing Conversion Drivers and Revenue Impact

(STAT543: Data Analysis Using Statistical Packages)

Shreya Rup Roy
20384316

Table of Contents

A/B Testing for Business Growth: Analysing Conversion Drivers and Revenue Impact

1. Introduction

The project evaluates how different landing page designs affect the user conversions using A/B testing. It further examines the role of demographics, traffic sources, user engagement, coupon usage and payment methods in influencing the conversion rates and revenue. And by analysing these factors, this project aims to provide data-driven insights to optimize website performance, marketing strategies and revenue generation.

2. Objective

This project analyses the impact of different landing page variants on user conversions while exploring the influence of demographics, traffic sources, engagement metrics, coupons, and payment methods on conversion rates and revenue.

The main objective is divided into sub-objectives which are listed below:

- **Compare conversion rates across different landing page variants.**
- **Analyse how demographics influence conversions.**
- **Examine the impact of traffic sources on user engagement.**
- **Investigate how time spent and pages visited affect conversions.**
- **Evaluate the effectiveness of coupons.**
- **Assess revenue impact based on payment methods.**

3. Methodology

The methodologies for each objective are given below:

❖ For **Objective 1:** Compare conversion rates across different landing page variants

Data Preparation:

- Grouped data by **variant_group** (Vibrant, Cold, Heat) to compare landing page performance.
- Calculated **total users, total conversions, and conversion rates** for each variant.

Statistical Testing (Chi-Square Test):

- Created a **contingency table** of conversions (1) and non-conversions (0) by variant.
- Performed a **Chi-Square test** to check if conversion rates differ significantly.
- **Hypotheses:**
  - **H_0:** Conversion rates are independent of the landing page.
  - **H_1:** Conversion rates differ across variants.

Visualization:

- Used a **bar chart** to visually compare conversion rates across the three landing page variants.

- For **Objective 2:** Analyse how demographics influence conversions
  - Grouped data by each of the demographic categories (like Age group, Gender, Location and Device type) and calculated their respective conversion rates.
  - Conducted a Chi-Square test for each of them to assess if conversion rates differ significantly across categories.
  - For location, chi-square test was not performed since it had 26 categories, which will make it difficult to show. So, instead I have showed the top 10 locations with the highest conversion rates.
  - Visualized the results using bar plots using ggplot2 and also highlighted trends and patterns for easy interpretation.

- For **Objective 3:** Examine the impact of traffic sources on user engagement
  - Grouped the data by traffic source then compute conversion rate and also calculate average time spent on the platform for each traffic source.
  - Create a contingency table for traffic source vs. conversion flag and then conduct a Chi-Square test to check if conversion rates significantly differ across traffic sources.
  - Perform ANOVA to determine if the average time spent differs significantly across different traffic sources.
  - Visualize the findings through Bar plots for both conversion rate and average time spent for each traffic sources.

- For **Objective 4:** Investigate how time spent and pages visited affect conversions
  - Built a logistic regression model to predict conversion probability using time spent and pages visited as predictor variables and examined the model summary to understand how these factors impact conversion.
  - Visualized user behaviour through scatter plot by plotting time spent vs. pages visited with points coloured based on conversion status and also added a trendline to observe the relationship between these factors.
  - Comparing Converters vs. Non-Converters through box plots for both time spent and pages visited.

- For **Objective 5:** Evaluate the effectiveness of coupons
  - Grouped data by coupon applied (Yes/No) and calculated total conversions, total users, and conversion rate for each group.
  - Performed a Chi-square test to see if the difference in conversion rates between coupon users and non-users is statistically significant.

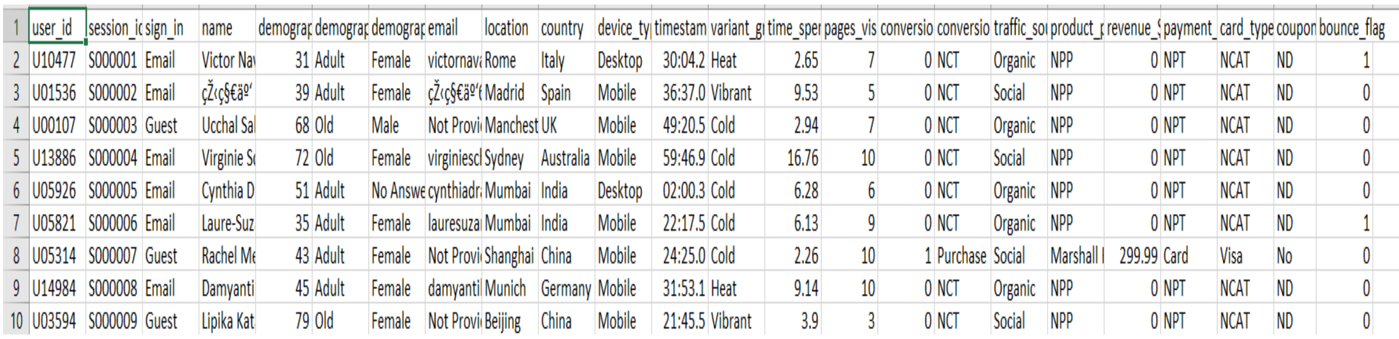
- Built a logistic regression model to assess whether applying a coupon increases the probability of conversion and examined model coefficients to understand the impact.
- Calculated total revenue and average revenue per user for both groups and analysed whether offering coupons leads to higher overall revenue.
- Created a bar chart to compare conversion rates for users who applied coupons vs. those who didn't.

❖ For **<u>Objective 6:</u>** Assess revenue impact based on payment methods

- Grouped data by payment method and calculated total revenue and average revenue per user for each payment type.
- Conducted an ANOVA to check if the average revenue significantly differs across different payment types.
- Filtered the data to focus only on Card payments and grouped by card type (e.g., Visa, Mastercard, etc.) to analyse total and average revenue per card type.
- Performed another ANOVA test to see if revenue varies significantly across different card types.
- Created two bar charts, one showing average revenue per payment type and another to compare average revenue per card type among Card users.

4. Overview of Dataset

The dataset used for the project is of Bluetooth Speaker sales dataset available in Kaggle which comprises of 30,000 user sessions, including both returning and first-time visitors. It is structured around different landing page variants users interacted with, enabling a comprehensive A/B testing analysis. The dataset captures key attributes such as user demographics, session engagement, product purchases, payment methods, and conversion outcomes, offering valuable insights into user behaviour and revenue trends.



1	user_id	session_ic	sign_in	name	demograp	demograp	demograp	email	location	country	device_ty	timestam	variant_g	time_sper	pages_vis	conversio	conversio	traffic_so	product_p	revenue_$	payment_	card_type	coupon	bounce_flag
2	U10477	S000001	Email	Victor Na	31	Adult	Female	victornav	Rome	Italy	Desktop	30:04.2	Heat	2.65	7	0	NCT	Organic	NPP	0	NPT	NCAT	ND	1
3	U01536	S000002	Email	çŽ‹ç§€äº'	39	Adult	Female	çŽ‹ç§€äº'(	Madrid	Spain	Mobile	36:37.0	Vibrant	9.53	5	0	NCT	Social	NPP	0	NPT	NCAT	ND	0
4	U00107	S000003	Guest	Ucchal Sa	68	Old	Male	Not Provi	Manchest	UK	Mobile	49:20.5	Cold	2.94	7	0	NCT	Organic	NPP	0	NPT	NCAT	ND	0
5	U13886	S000004	Email	Virginie S	72	Old	Female	virginiescl	Sydney	Australia	Mobile	59:46.9	Cold	16.76	10	0	NCT	Social	NPP	0	NPT	NCAT	ND	0
6	U05926	S000005	Email	Cynthia D	51	Adult	No Answe	cynthiadr	Mumbai	India	Desktop	02:00.3	Cold	6.28	6	0	NCT	Organic	NPP	0	NPT	NCAT	ND	0
7	U05821	S000006	Email	Laure-Suz	35	Adult	Female	lauresuza	Mumbai	India	Mobile	22:17.5	Cold	6.13	9	0	NCT	Organic	NPP	0	NPT	NCAT	ND	1
8	U05314	S000007	Guest	Rachel Me	43	Adult	Female	Not Provi	Shanghai	China	Mobile	24:25.0	Cold	2.26	10	1	Purchase	Social	Marshall	299.99	Card	Visa	No	0
9	U14984	S000008	Email	Damyanti	45	Adult	Female	damyanti	Munich	Germany	Mobile	31:53.1	Heat	9.14	10	0	NCT	Organic	NPP	0	NPT	NCAT	ND	0
10	U03594	S000009	Guest	Lipika Kat	79	Old	Female	Not Provi	Beijing	China	Mobile	21:45.5	Vibrant	3.9	3	0	NCT	Social	NPP	0	NPT	NCAT	ND	0

<u>Note:</u> This image shoes the first 10 rows of the dataset.

The key attributes that were used for the project are:

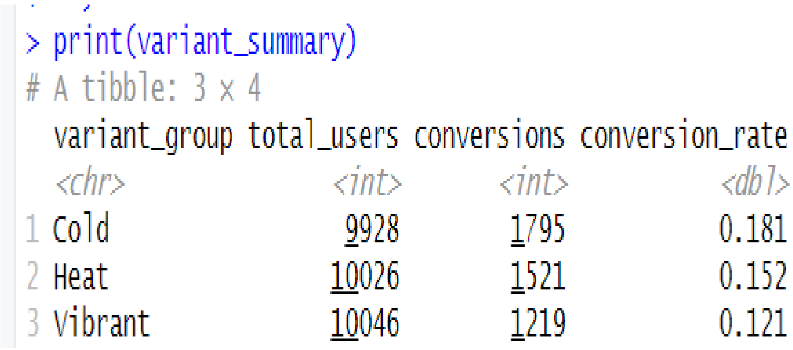
a) demographic_age_group: Age group of the visitor (Adult, Teenage, Old)
b) demographic_gender: Gender of the visitor (Male, Female, Not Answered)
c) location: The city where the visitor is located

d) device_type: Type of device used by the visitor (Mobile, Tablet, Desktop)
e) variant_group: The landing page design variant the user saw (Vibrant, Cold, Heat)
f) time_spent: Total time (in minutes) the user spent on the landing page
g) pages_visited: Number of pages the user viewed during the session
h) conversion_flag: Binary flag (0/1) indicating whether the user converted (i.e. signed up or made a purchase)
i) traffic_source: Tells sources of traffic (Organic: visitors who arrive at the website through unpaid search engine results; Paid: visitors who land on website through paid advertising campaigns; Social: visitors coming from social media platforms; Referral: users who visit the website by clicking on links from other websites, blogs, etc.)
j) revenue: Total revenue generated from the transaction (if purchase was made)
k) payment_type: Payment methods used by the users (Card, COD, Not Provided Type)
l) card_type: Card type if payment was made by card (Amex, Visa, Master)
m) coupon_applied: tells whether a coupon was applied (Yes, No, Not Determined)

The project was carried out using R.

5. Results and Discussions

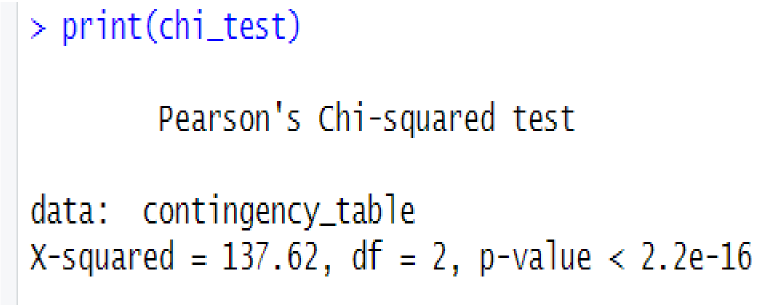
❖ Objective 1 result shows:



```
> print(variant_summary)
# A tibble: 3 x 4
  variant_group total_users conversions conversion_rate
  <chr>               <int>       <int>           <dbl>
1 Cold                 9928        1795           0.181
2 Heat                10026        1521           0.152
3 Vibrant             10046        1219           0.121
```

Interpretation:

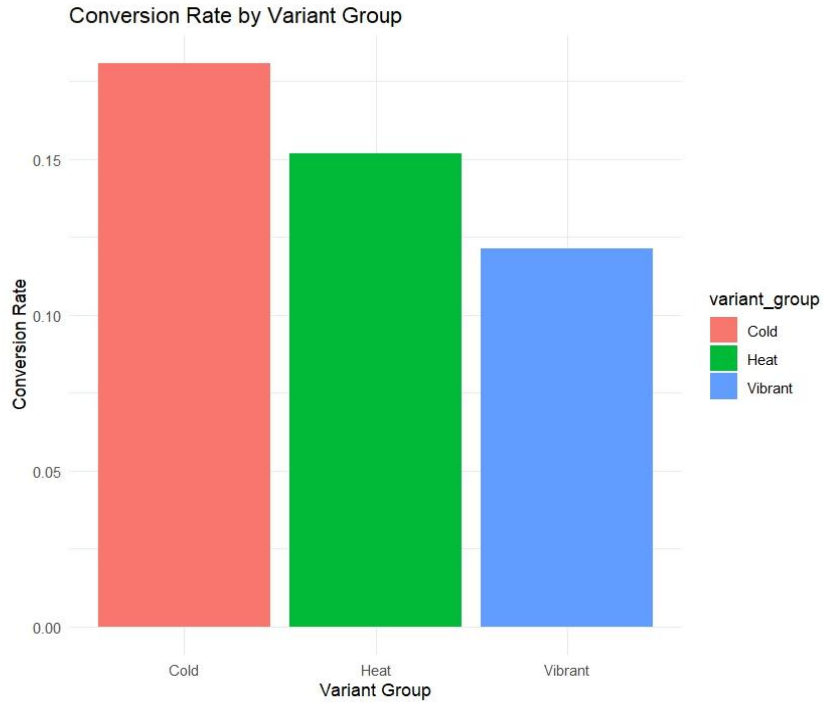
- The Cold landing page variant had the highest conversion rate at 18.1%, followed by Heat variant at 15.2% and Vibrant variant had the lowest conversion rate at 12.1%.



```
> print(chi_test)

        Pearson's Chi-squared test

data:  contingency_table
X-squared = 137.62, df = 2, p-value < 2.2e-16
```

Interpretation:

- The p-value is very small, so we reject the null hypothesis, meaning at least one of the variants performed differently from the others in terms of conversion rates.



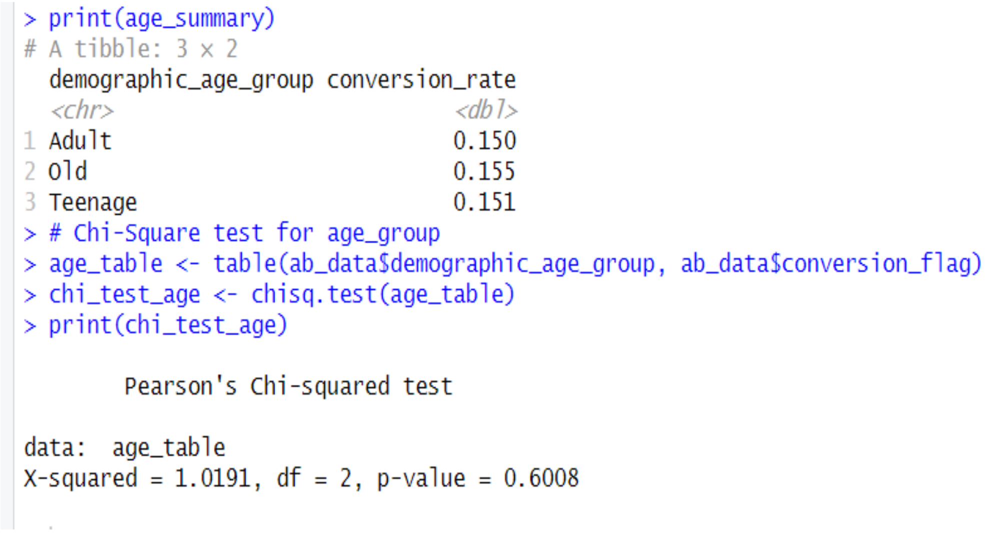
__Interpretation:__ This bar plot shows that the Cold variant had the highest conversion rate, followed by Heat variant and Vibrant variant being the least one among them.

Business Recommendation for Objective 1:

- ❑ **If the goal is to maximize conversions, the Cold variant should be prioritized or further analysed for improvement since the Cold variant appears to be the most effective in driving conversions.**
- ❑ **Modifications can be made on the Heat and Vibrant variants to optimize conversion performance.**

❖ Objective 2 result shows:

1) Conversion rates by demographic age group and chi-square test for the age group



```
> print(age_summary)
# A tibble: 3 x 2
  demographic_age_group conversion_rate
  <chr>                           <dbl>
1 Adult                           0.150
2 Old                             0.155
3 Teenage                         0.151
> # Chi-Square test for age_group
> age_table <- table(ab_data$demographic_age_group, ab_data$conversion_flag)
> chi_test_age <- chisq.test(age_table)
> print(chi_test_age)

        Pearson's Chi-squared test

data:  age_table
X-squared = 1.0191, df = 2, p-value = 0.6008
```

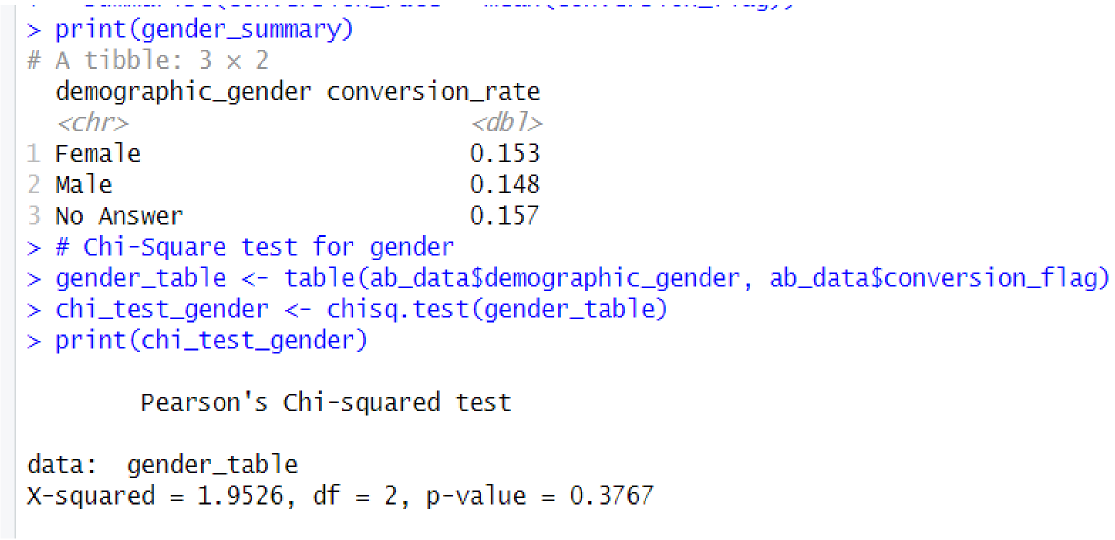
Interpretation:

- The Old age group had the highest conversion rates at 15.5%, followed by Teenage age group at 15.1% and Adult age group had the lowest conversion rate at 15.1%. The conversion rates across age groups are quite similar, with minor differences.
- Since the p-value (0.6008) is greater than 0.05, we fail to reject the null hypothesis. This means that there is no statistically significant difference in conversion rates across different age groups.

Business Recommendation for Objective 2.1:

❑ **Since age does not significantly influence conversions, marketing efforts should focus on other factors, such as user behaviour, engagement metrics, or other demographic features.**

2) Conversion rates by demographic gender and chi-square test for the gender



```
> print(gender_summary)
# A tibble: 3 × 2
  demographic_gender conversion_rate
  <chr>                        <dbl>
1 Female                       0.153
2 Male                         0.148
3 No Answer                    0.157
> # Chi-Square test for gender
> gender_table <- table(ab_data$demographic_gender, ab_data$conversion_flag)
> chi_test_gender <- chisq.test(gender_table)
> print(chi_test_gender)

	Pearson's Chi-squared test

data:  gender_table
X-squared = 1.9526, df = 2, p-value = 0.3767
```

Interpretation:

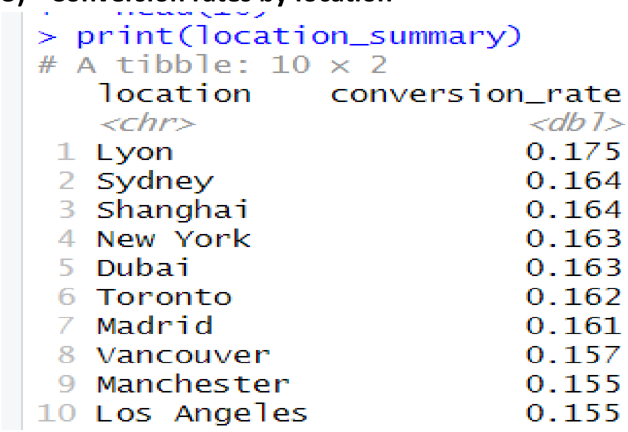
- No Answer (i.e. the users who had not specified their gender) had the highest conversion rates at 15.7%, followed by Females at 15.3%, and Males had the lowest conversion rates at 14.8%. The differences in conversion rates among gender groups are small.
- Since the p-value (0.3767) is greater than 0.05, we fail to reject the null hypothesis. This shows there is no strong evidence to suggest that gender significantly affects conversion rates in this dataset.

Business Recommendation for Objective 2.2:

❑ **Since there is no strong gender-based difference in conversion rates, businesses should focus on gender-neutral branding and messaging to appeal to a broader audience.**

❑ **Investment in personalized content or targeted advertising should be based on other variables like location, or browsing behaviour.**

3) Conversion rates by location



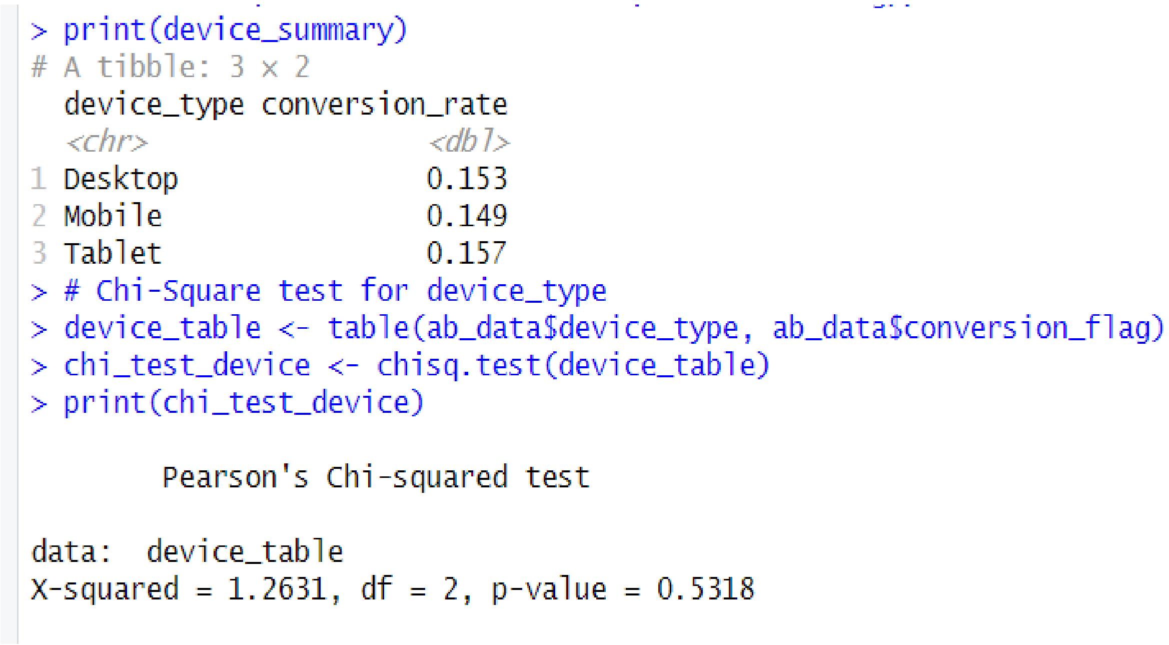
```
> print(location_summary)
# A tibble: 10 × 2
   location      conversion_rate
   <chr>                   <dbl>
 1 Lyon                    0.175
 2 Sydney                  0.164
 3 Shanghai                0.164
 4 New York                0.163
 5 Dubai                   0.163
 6 Toronto                 0.162
 7 Madrid                  0.161
 8 Vancouver               0.157
 9 Manchester              0.155
10 Los Angeles             0.155
```

Interpretation: This result shows the top 10 locations with the highest conversion rates with Lyon (in France) being at the top.

Business Recommendation for Objective 2.3:

- ❑ **Allocate more resources, advertising budget, and localized campaigns to the cities with higher conversion rates to further enhance engagement and conversions.**
- ❑ **Evaluate current marketing strategies and optimize ad placements, pricing, and promotions in mid-tier locations to push conversions higher.**
- ❑ **Investigate potential barriers to conversion in underperforming cities, such as pricing, competition, or customer preferences to overcome the problem.**
- ❑ **There may be market expansion opportunities in regions performing well.**

4) Conversion rates by device type and chi-square test for the device type



```
> print(device_summary)
# A tibble: 3 × 2
  device_type conversion_rate
  <chr>                 <dbl>
1 Desktop               0.153
2 Mobile                0.149
3 Tablet                0.157
> # Chi-Square test for device_type
> device_table <- table(ab_data$device_type, ab_data$conversion_flag)
> chi_test_device <- chisq.test(device_table)
> print(chi_test_device)

	Pearson's Chi-squared test

data:  device_table
X-squared = 1.2631, df = 2, p-value = 0.5318
```

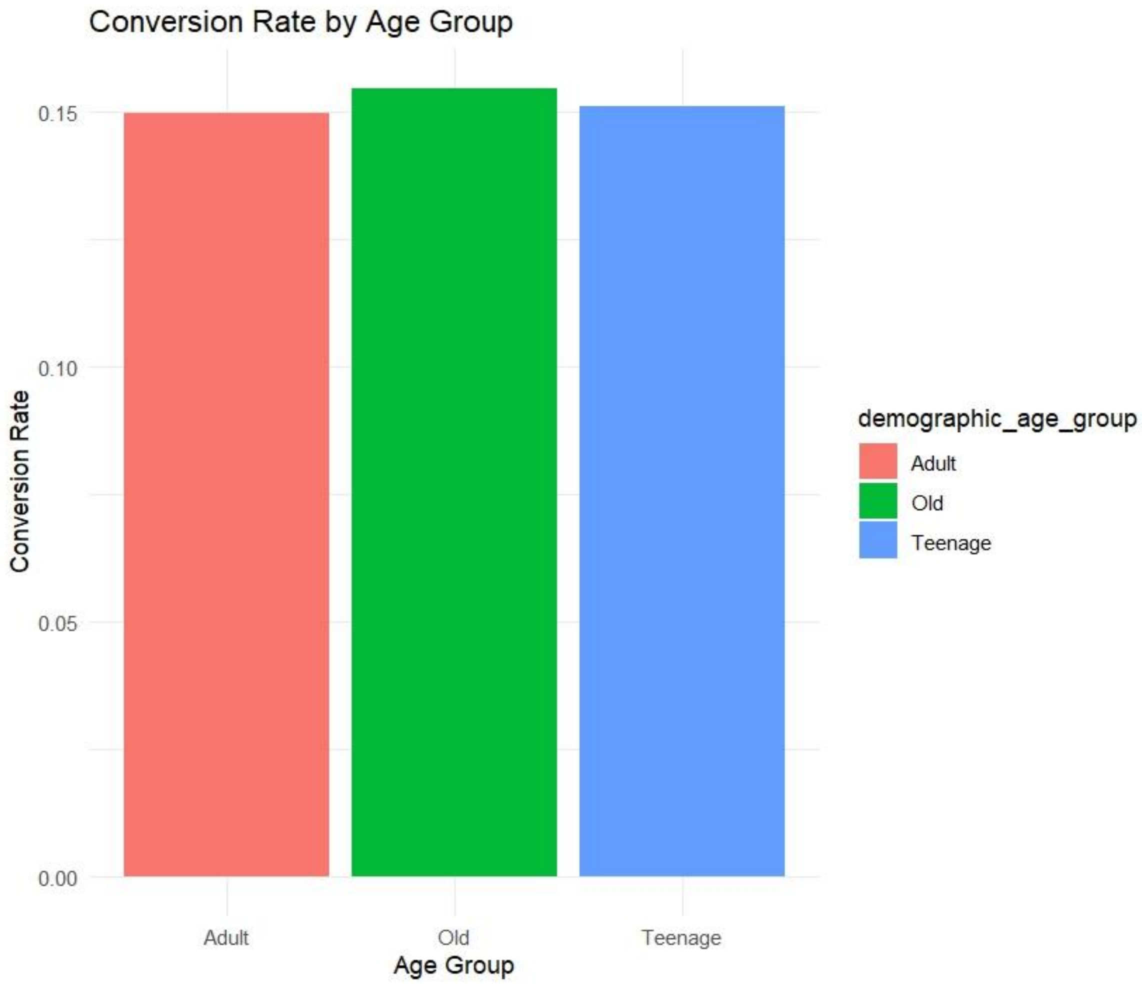
Interpretation:

- The device type Tablet had the highest conversion rates at 15.7%, followed by Desktop at 15.3% and Mobile had the lowest conversion rates at 14.9%. The differences in conversion rates among devices are small.
- Since the p-value (0.5318) is greater than 0.05, we fail to reject the null hypothesis. This suggests there is no statistically significant difference in conversion rates across devices.

Business Recommendation for Objective 2.4:

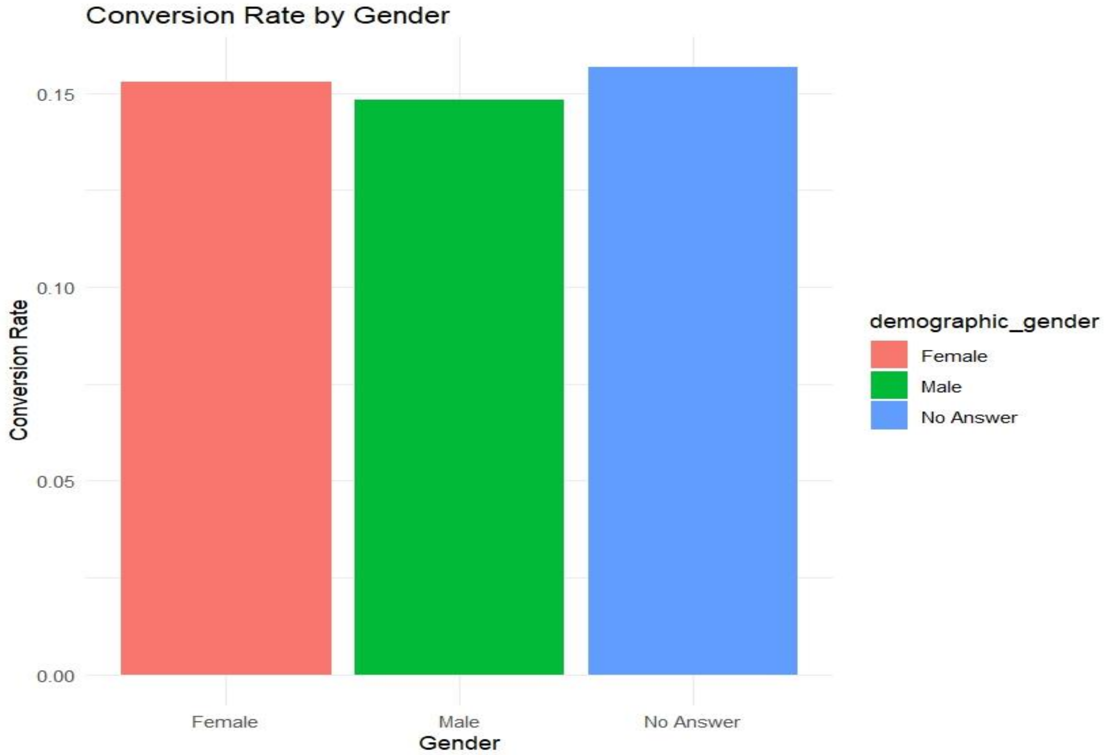
- **Since device type does not significantly impact conversion rates, the business should focus on overall user experience and content strategy rather than device-specific optimizations.**
- **However, mobile experience improvements could still be explored to maximize performance.**

5) Visualization for age group



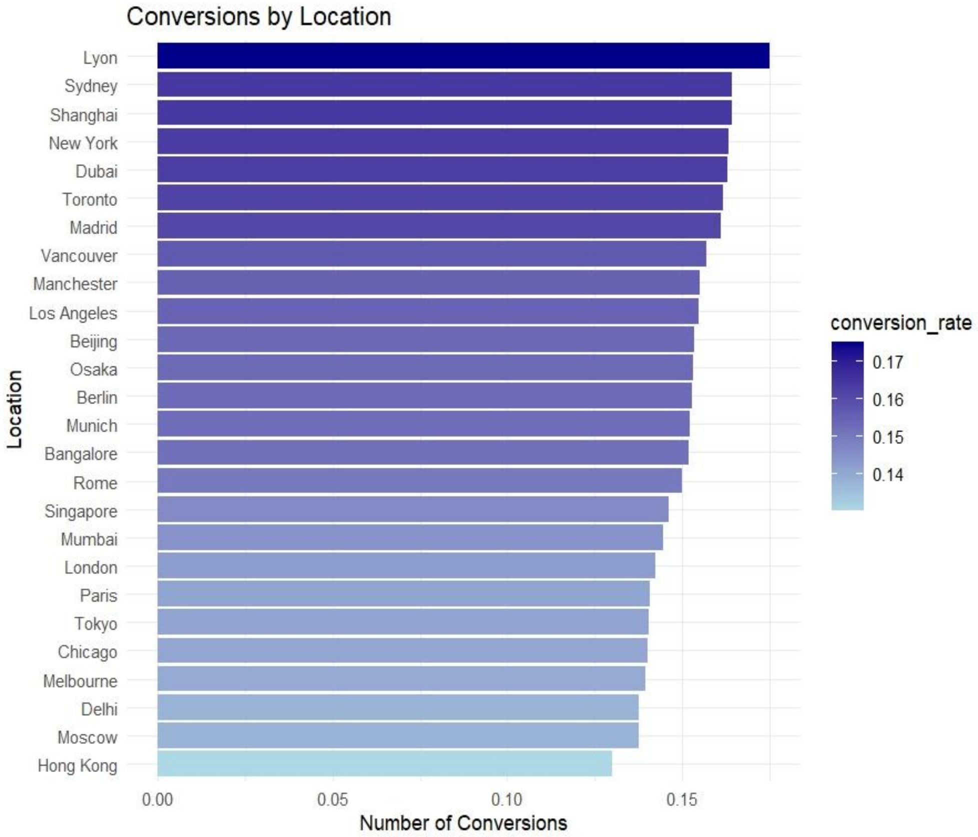
Interpretation: This bar plot shows that the age group Old had the highest conversion rates while Adult observed the lowest. The result is same as the chi-square test.

6) Visualization by gender



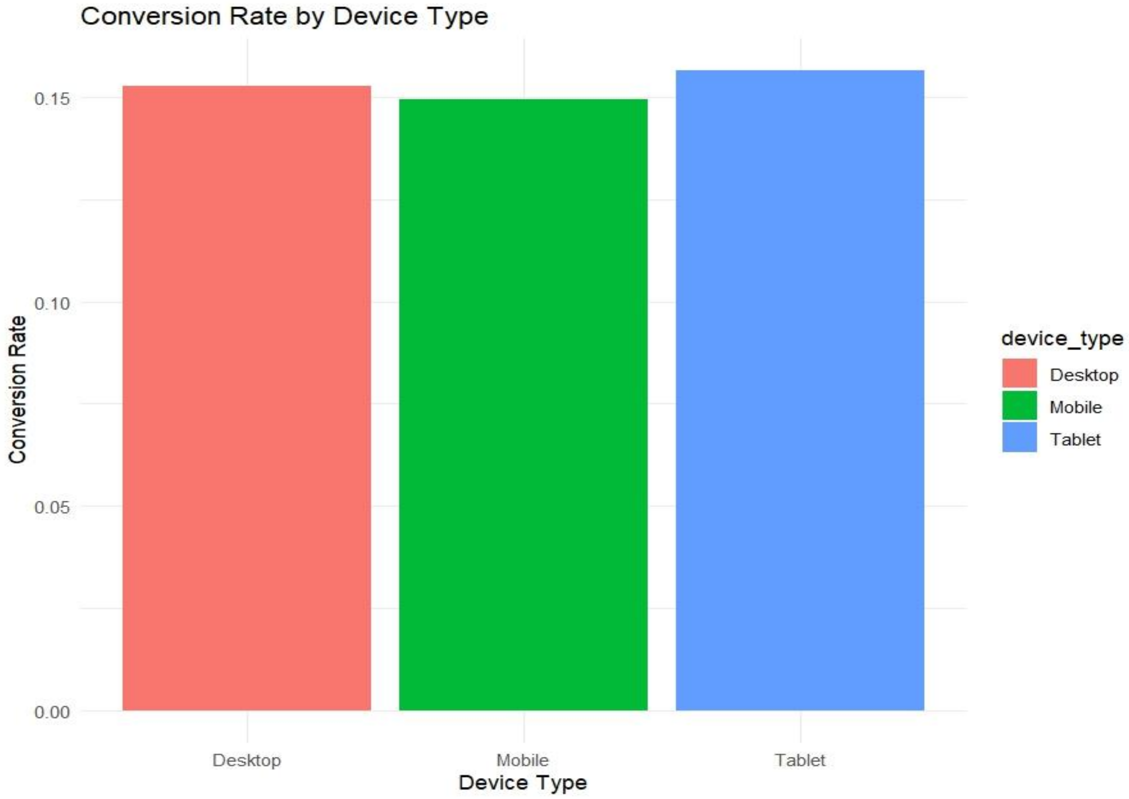
Interpretation: This bar plot shows the conversion rates for the No Answer category had the highest conversion rates while Males had the lowest.

7) Visualization by location



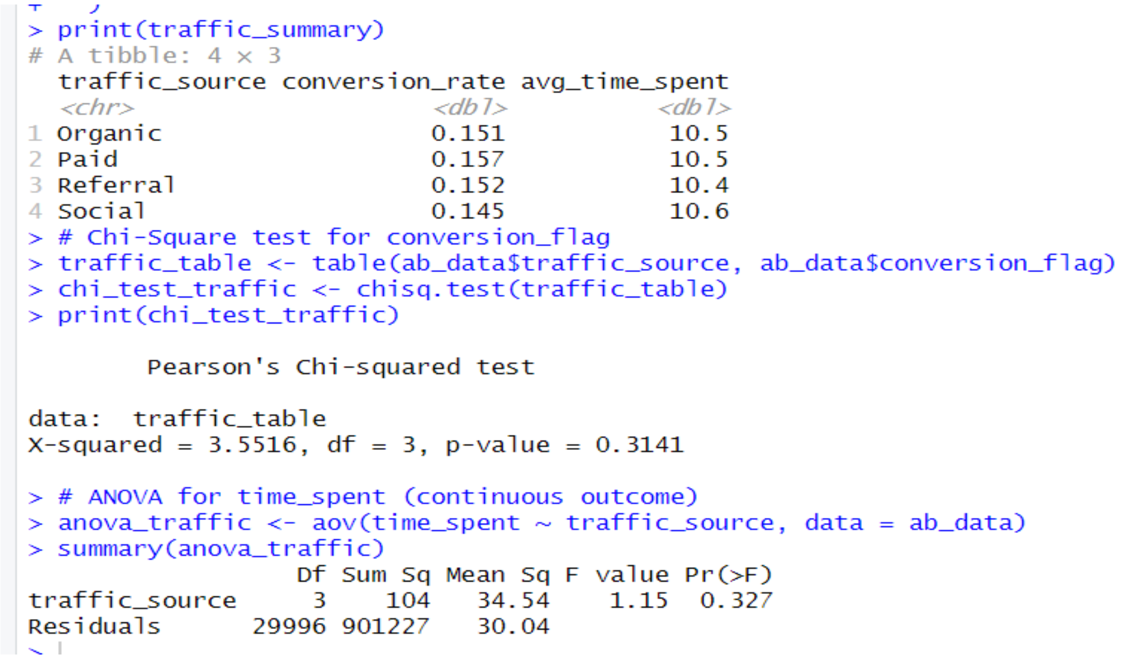
Interpretation: The top 3 locations in terms of conversion rates are Lyon, Sydney and Shanghai. While, the least 3 in terms of conversions are Delhi, Moscow and Hong Kong.

8) Visualization by device type



Interpretation: Tablet had the highest conversion rates while Desktop had the lowest among the devices.

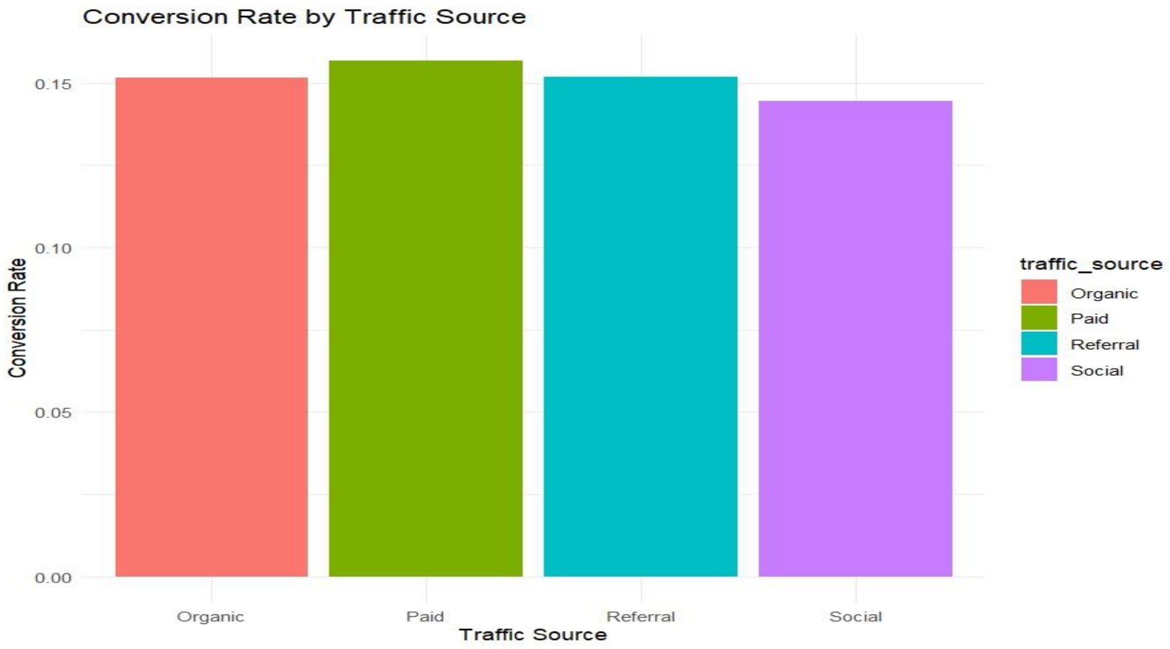
- Objective 3 result shows:



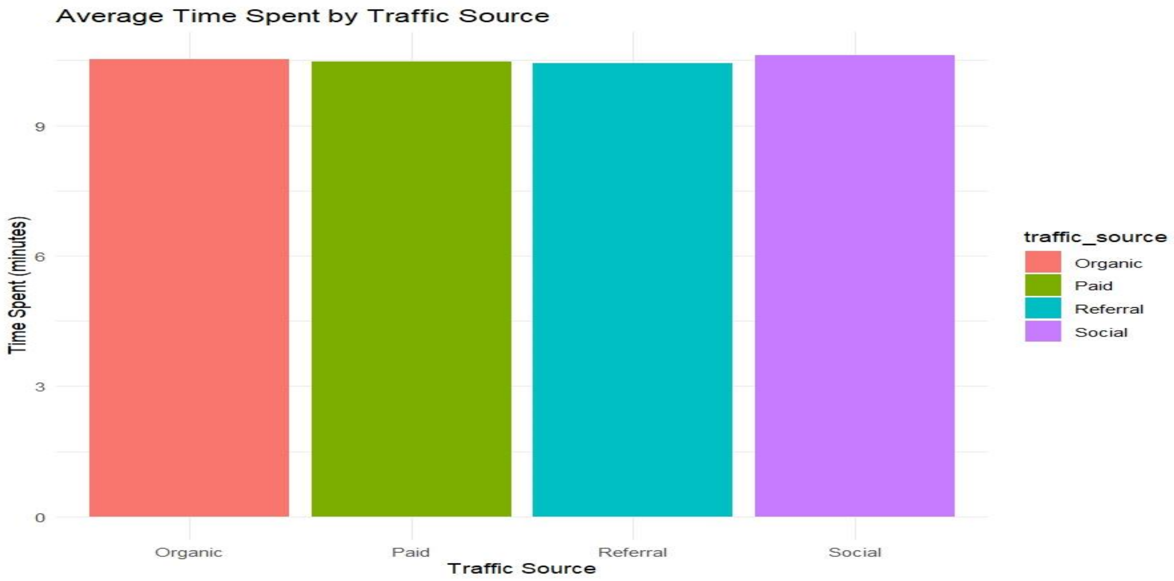
```
> print(traffic_summary)
# A tibble: 4 × 3
  traffic_source conversion_rate avg_time_spent
  <chr>                    <dbl>          <dbl>
1 Organic                  0.151           10.5
2 Paid                     0.157           10.5
3 Referral                 0.152           10.4
4 Social                   0.145           10.6
> # Chi-Square test for conversion_flag
> traffic_table <- table(ab_data$traffic_source, ab_data$conversion_flag)
> chi_test_traffic <- chisq.test(traffic_table)
> print(chi_test_traffic)

        Pearson's Chi-squared test

data:  traffic_table
X-squared = 3.5516, df = 3, p-value = 0.3141

> # ANOVA for time_spent (continuous outcome)
> anova_traffic <- aov(time_spent ~ traffic_source, data = ab_data)
> summary(anova_traffic)
                  Df Sum Sq Mean Sq F value Pr(>F)
traffic_source     3    104   34.54    1.15  0.327
Residuals      29996 901227   30.04
```

Interpretation:

- The conversion rates are pretty similar across all traffic sources, with Paid traffic being slightly higher and Social traffic being slightly lower and people from all traffic sources spend roughly the same amount of time on the website on average.
- The p-value (0.3141) is greater than the common threshold of 0.05. This means that we don't have enough statistical evidence to say that the traffic source significantly affects whether someone converts or not.
- Similar to the conversion test, the p-value (0.327) is greater than 0.05. This means that we don't have enough statistical evidence to say that the traffic source significantly affects how much time people spend on the website.



Interpretation: Paid traffic had the highest conversion rates while social traffic had the lowest among the various traffic sources.



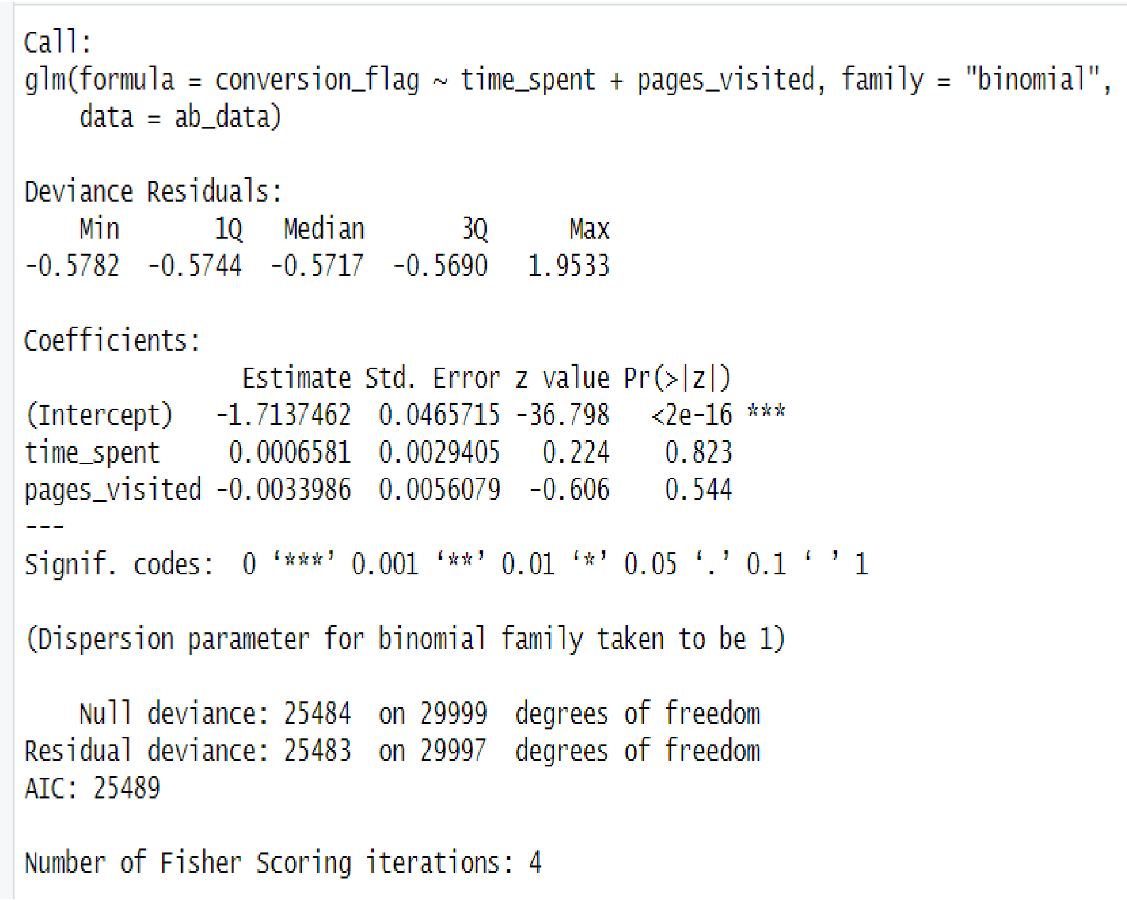
Interpretation: People from different sources spend about the same amount of time on the website, on average with Social traffic being just slightly higher.

Business Recommendation for Objective 3:

- ❑ **The current data suggests a balanced performance across your traffic sources. The primary focus should likely be on optimizing the overall website experience to improve conversions and engagement for all users.**
- ❑ **Deeper qualitative analysis within each channel can provide further insights for targeted improvements within specific traffic sources.**

❖ Objective 4 result shows:

1) Logistic regression for conversion_flag

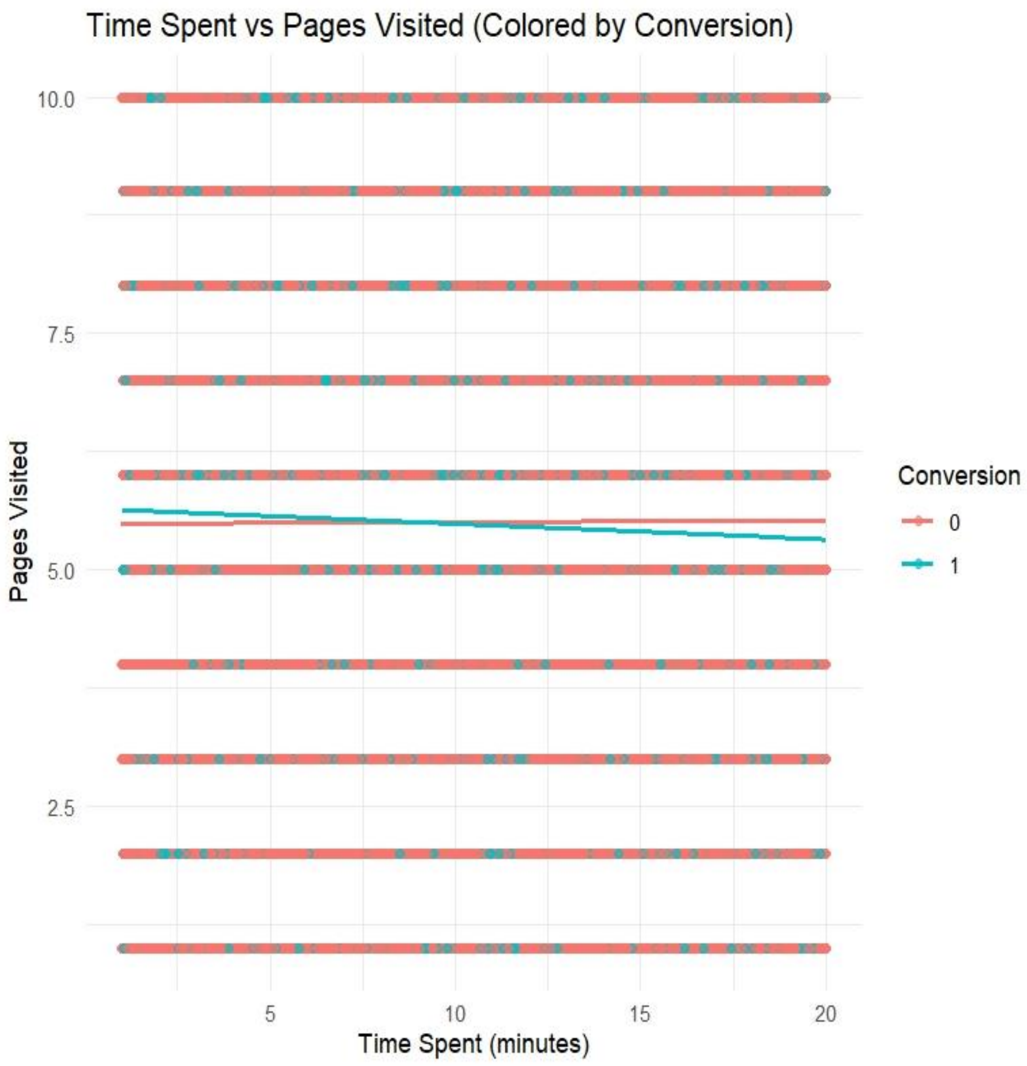


```
Call:
glm(formula = conversion_flag ~ time_spent + pages_visited, family = "binomial",
    data = ab_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5782  -0.5744  -0.5717  -0.5690   1.9533

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.7137462  0.0465715 -36.798   <2e-16 ***
time_spent     0.0006581  0.0029405   0.224    0.823
pages_visited -0.0033986  0.0056079  -0.606    0.544
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25484  on 29999  degrees of freedom
Residual deviance: 25483  on 29997  degrees of freedom
AIC: 25489

Number of Fisher Scoring iterations: 4
```

Interpretation:

- The p-value for time_spent is (0.823) which is not significant (p > 0.05), meaning there's no strong evidence that time_spent has a meaningful effect on conversion in this model.
- Similarly, the p-value for pages_visited (0.544) is also not significant (p > 0.05), suggesting that pages_visited does not have a statistically significant effect on conversion in this model.
- The difference between null and residual deviance (25484 - 25483 = 1) is very small, indicating that the predictors time_spent and pages_visited do not explain much of the variability in conversion_flag.

2) How users who converted behave in terms of time spent and pages visited?

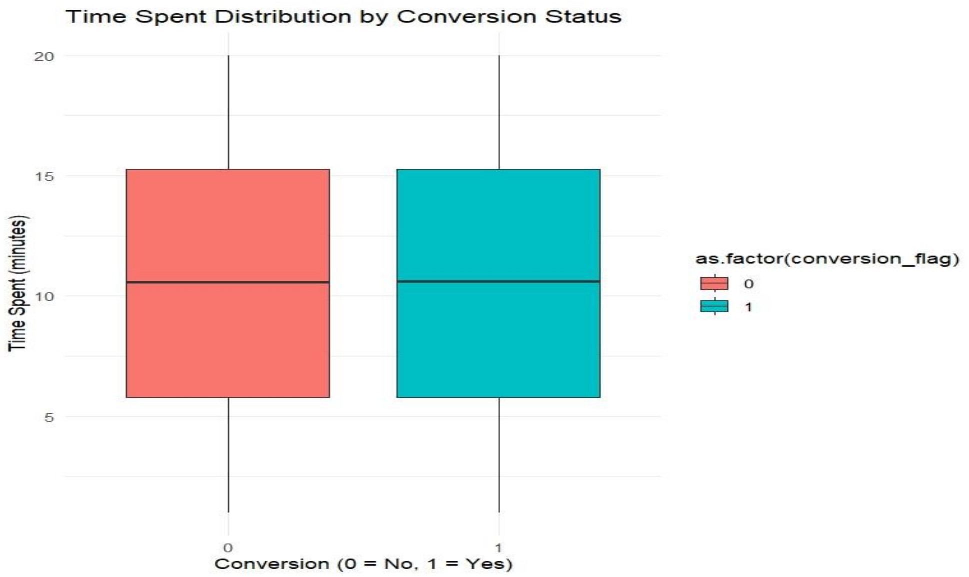


Interpretation:

- Most Sessions are Short: A large number of data points are clustered towards the lower end of the "Time Spent" axis (closer to 0 minutes).
- Discrete Page Visits: The "Pages Visited" data appears to be discrete, with clusters of points at whole or half-integer values (e.g., 1, 2, 2.5, 3, etc.). This suggests that the number of pages visited is likely a count or a value with limited granularity.
- Conversion Distribution: Conversions (teal dots) appear to be scattered across different levels of "Pages Visited" and "Time Spent." There isn't an immediately obvious strong visual pattern suggesting that conversions are heavily concentrated in specific ranges of time spent or pages visited.
- Trend lines suggest that:
  - For people who *don't* convert, spending more time on the site doesn't really mean they look at more pages on average.
  - For people who *do* convert, there's a slight tendency to see a few more pages if they spend a bit more time, but even that levels off.

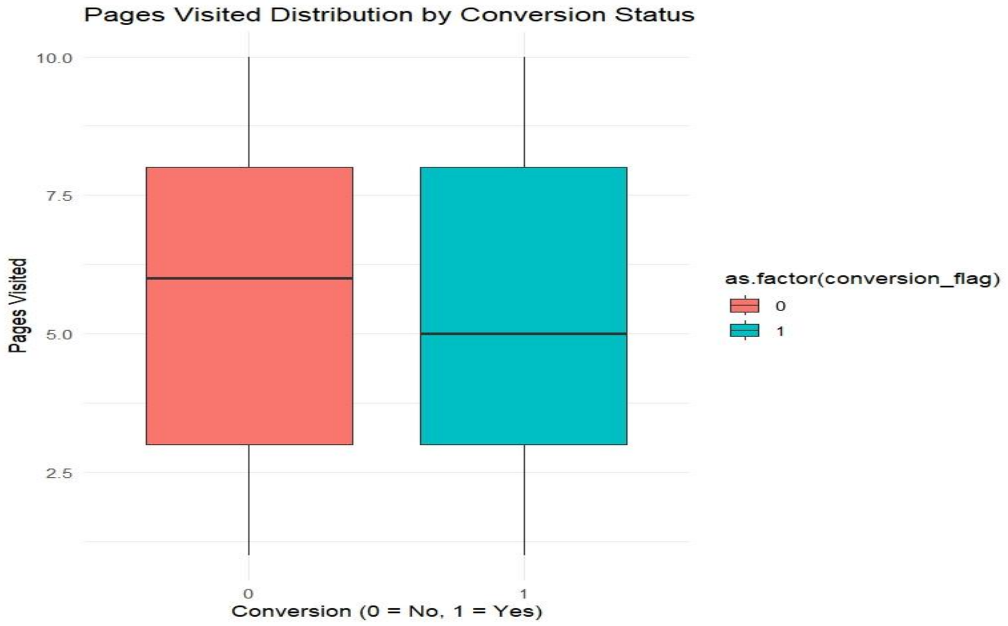
3) Boxplots to compare converters vs non-converters

a) Boxplot for time spent by conversions



Interpretation: There is no strong visible difference in time spent between converted and non-converted users.

b) Boxplot for pages visited by conversions



Interpretation:

- People who didn't convert tended to look at a slightly higher number of pages on average compared to those who did convert.
- However, the typical range of pages visited for both groups is quite similar. Most people, whether they converted or not, visited somewhere between 3 and 8 pages.
- Overall, the number of pages visited doesn't seem to be a very strong indicator of whether someone will convert or not, as the distributions are quite overlapping.

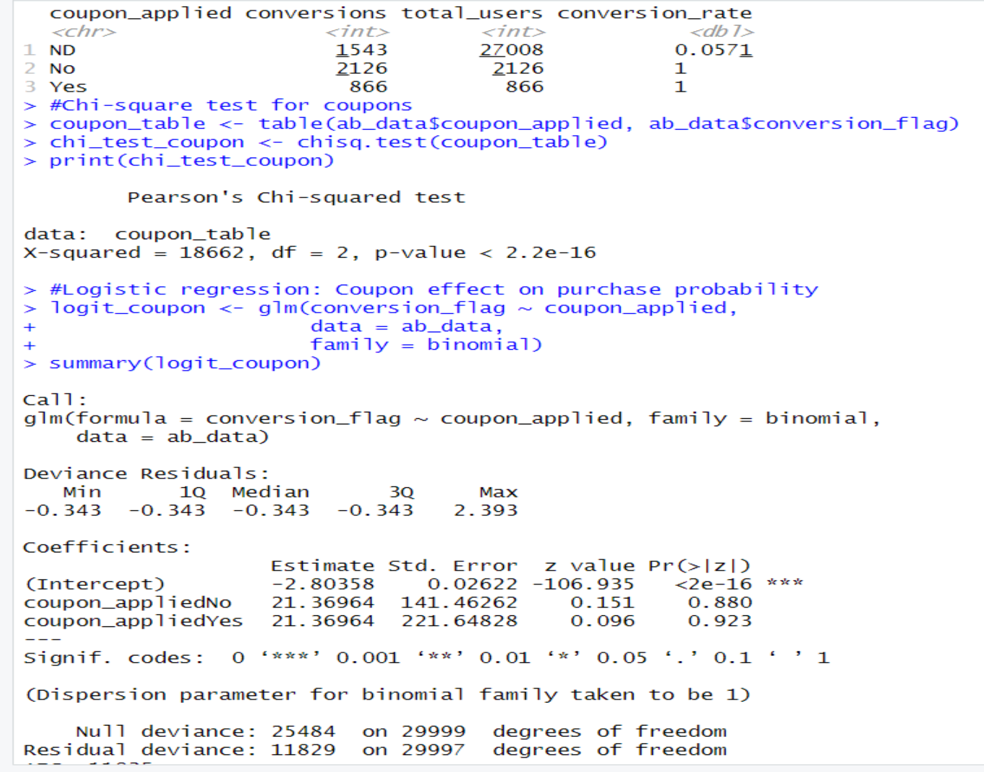
Business Recommendation for Objective 4:

❑ **To understand *why* time spent and page views aren't strong predictors, conduct qualitative research such as user interviews or usability testing. This can provide**

valuable insights into user behaviour and identify key drivers of conversion that quantitative data alone might miss.

❖ Objective 5 result shows:

1) **Compute conversion rate by coupon usage and perform Chi-square test, logistic regression to see if coupons had effect on purchases**

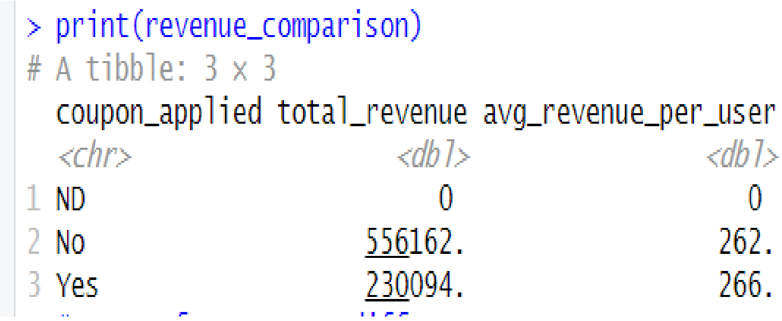


```
  coupon_applied conversions total_users conversion_rate
  <chr>                <int>       <int>           <dbl>
1 ND                    1543       27008          0.0571
2 No                    2126        2126          1
3 Yes                    866         866          1
> #Chi-square test for coupons
> coupon_table <- table(ab_data$coupon_applied, ab_data$conversion_flag)
> chi_test_coupon <- chisq.test(coupon_table)
> print(chi_test_coupon)

        Pearson's Chi-squared test

data:  coupon_table
X-squared = 18662, df = 2, p-value < 2.2e-16

> #Logistic regression: Coupon effect on purchase probability
> logit_coupon <- glm(conversion_flag ~ coupon_applied,
+                     data = ab_data,
+                     family = binomial)
> summary(logit_coupon)

Call:
glm(formula = conversion_flag ~ coupon_applied, family = binomial,
    data = ab_data)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-0.343  -0.343  -0.343  -0.343   2.393

Coefficients:
                   Estimate Std. Error  z value Pr(>|z|)
(Intercept)        -2.80358    0.02622 -106.935   <2e-16 ***
coupon_appliedNo   21.36964  141.46262    0.151    0.880
coupon_appliedYes  21.36964  221.64828    0.096    0.923
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25484  on 29999  degrees of freedom
Residual deviance: 11829  on 29997  degrees of freedom
```

Interpretation:

- The initial summary table shows unusual 100% conversion rates for the "Yes" and "No" coupon categories, but these are based on extremely small sample sizes (only 1 user each) and are likely not representative.
- The chi-square test strongly indicates that there *is* a relationship between coupon application and conversion overall since the p-value is less than 0.05.
- However, the logistic regression model, likely hampered by the unreliable data for "Yes" and "No," fails to provide statistically significant evidence of how specifically applying or not applying a coupon changes the likelihood of conversion compared to the "Not Determined" group.

2) **Compare revenue between coupon users and non-users**

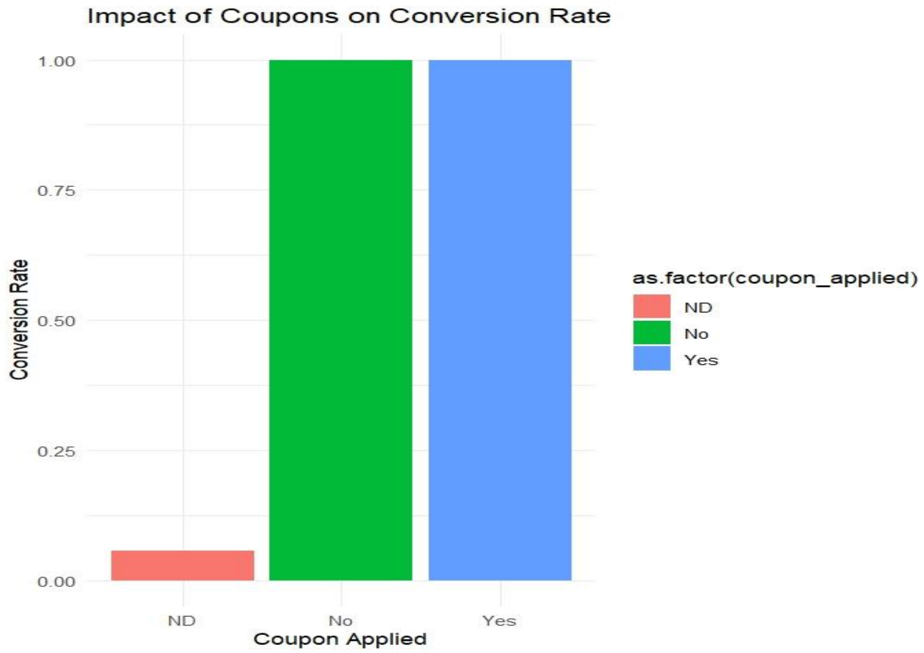


```
> print(revenue_comparison)
# A tibble: 3 x 3
  coupon_applied total_revenue avg_revenue_per_user
  <chr>                  <dbl>                <dbl>
1 ND                         0                    0
2 No                    556162.                 262.
3 Yes                   230094.                 266.
```

Interpretation:

- People who didn't use a coupon generated more total revenue of 556162.
- However, on average, the people who *did* use a coupon spent slightly more money per person compared to those who didn't use a coupon.

3) Plot coupon vs conversion rates



Interpretation: The conversion rates for both the groups who applied the coupons and not applied the coupons are 100% while the conversion rate for Not Determined category is very low.

Important Note: **This bar plot reflects those 100% conversion rates, but because the sample size is so tiny, these results might not be representative of the overall user behaviour. The "ND" group likely has a much larger number of users, making its low conversion rate more indicative of the general trend when coupon application isn't tracked.**

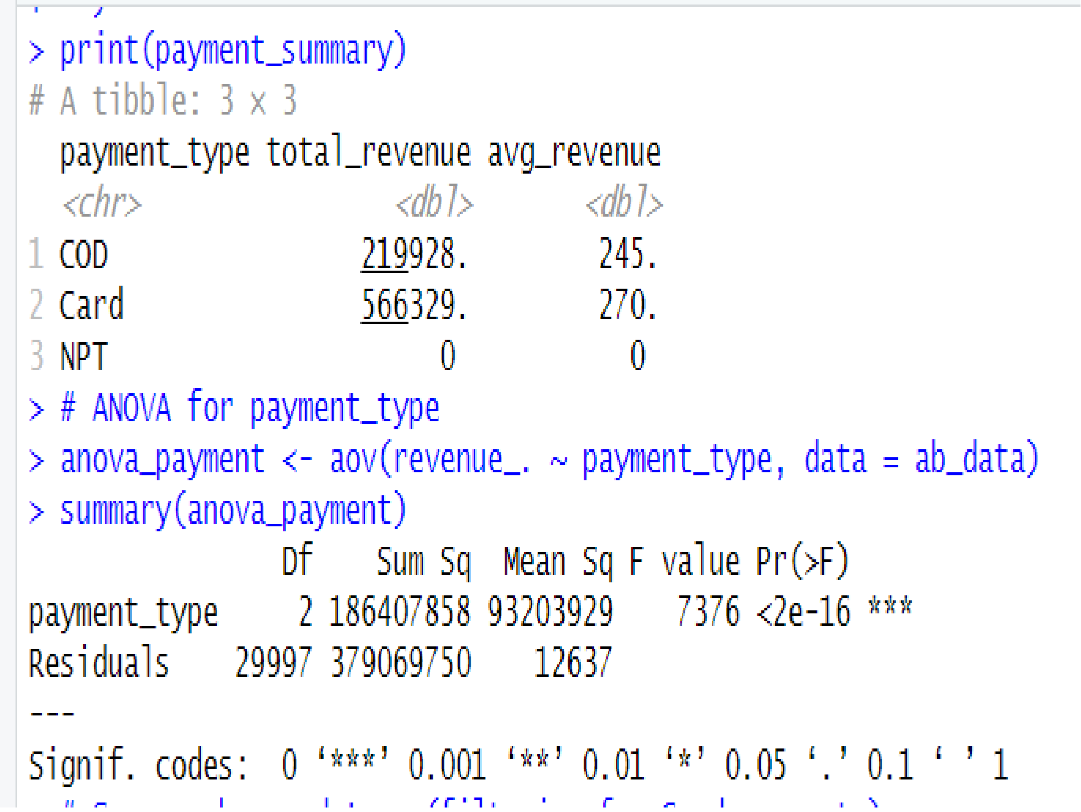
Business Recommendation for Objective 5:

Coupon Usage Likely Influences Conversion, But the Exact Nature is Unclear for which the following things need to be done:

- **Immediately improve coupon tracking mechanism.**
- **Gather more data on users who explicitly use or don’t use coupons.**
- **Consider A/B testing different coupon strategies.**
- **Analyse the customer journey for the “ND (Not Determined)” group.**

- ❖ Objective 6 result shows:

1) Summary by and ANOVA for payment type



```
> print(payment_summary)
# A tibble: 3 x 3
  payment_type total_revenue avg_revenue
  <chr>                <dbl>       <dbl>
1 COD                219928.        245.
2 Card               566329.        270.
3 NPT                     0           0
> # ANOVA for payment_type
> anova_payment <- aov(revenue_. ~ payment_type, data = ab_data)
> summary(anova_payment)
                Df    Sum Sq  Mean Sq F value Pr(>F)    
payment_type     2 186407858 93203929    7376 <2e-16 ***
Residuals    29997 379069750    12637                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

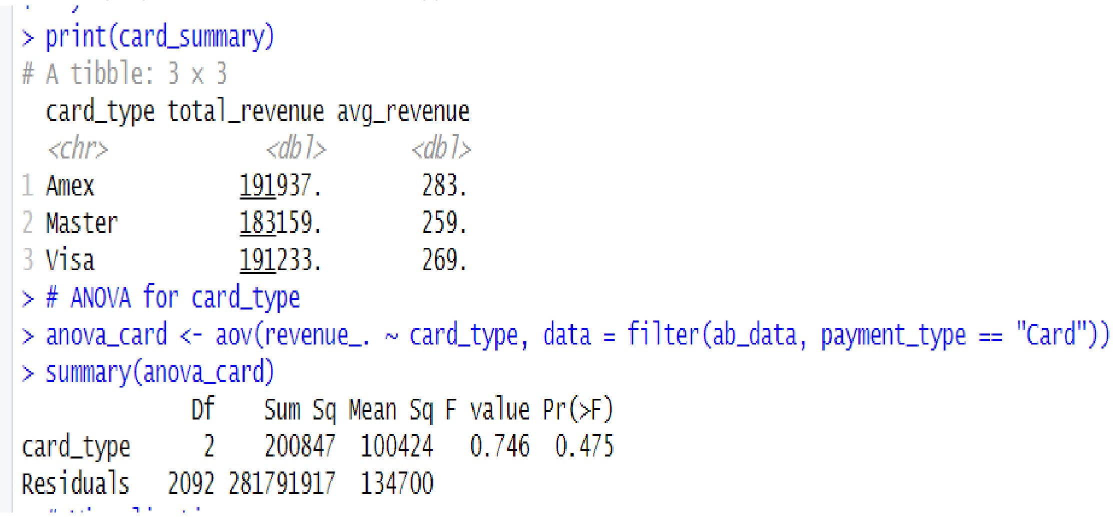
Interpretation:

- Most of the money came from people paying with cards.
- People who paid with cards also spent a bit more on average per transaction compared to those using Cash On Delivery.
- There are some transactions where the payment type isn't clear, and these didn't generate any revenue.
- The p-value < 0.05 which strongly indicates that there is a statistically significant difference in the average revenue generated by the different payment types. This means that the choice of payment method significantly affects the average revenue per transaction.

Business Recommendation for Objective 6.1:

- ❑ **Encourage Card Payments given the higher total and average revenue from card payments.**
- ❑ **While COD has a lower average revenue, it still contributes significantly, so optimize COD Process.**
- ❑ **Investigate "NPT", to find the issues that need to be addressed to recover potential revenue.**

2) Summary by and ANOVA for card type



```
> print(card_summary)
# A tibble: 3 x 3
  card_type total_revenue avg_revenue
  <chr>             <dbl>       <dbl>
1 Amex            191937.        283.
2 Master          183159.        259.
3 Visa            191233.        269.
> # ANOVA for card_type
> anova_card <- aov(revenue_. ~ card_type, data = filter(ab_data, payment_type == "Card"))
> summary(anova_card)
              Df    Sum Sq Mean Sq F value Pr(>F)
card_type      2    200847  100424   0.746  0.475
Residuals   2092 281791917  134700
```

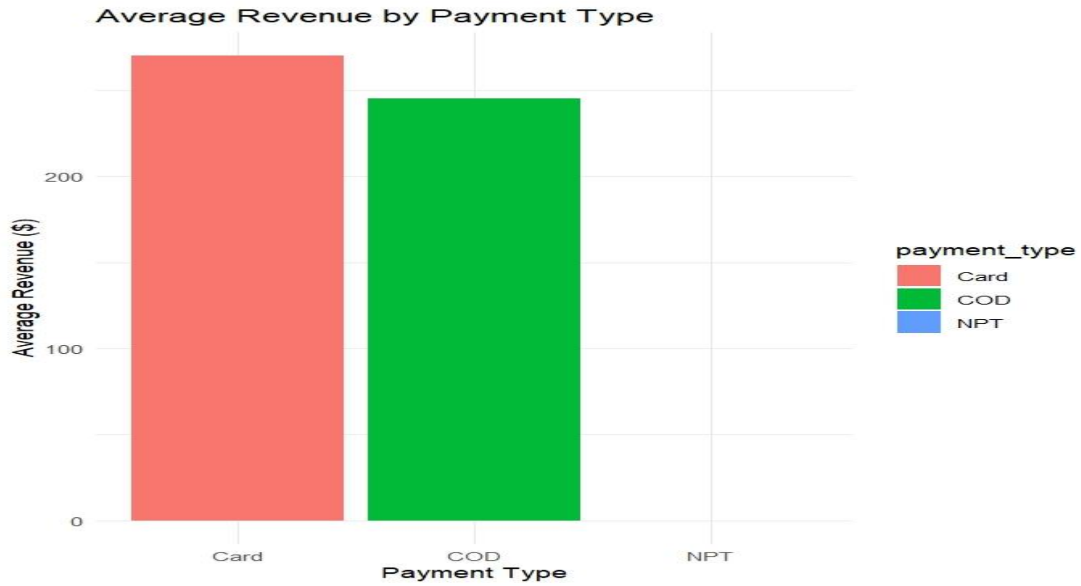
Interpretation:

- All three major card types (Amex, Mastercard, Visa) contributed roughly similar amounts to the total revenue from card payments.
- People using Amex tended to spend the most on average per transaction.
- People using Mastercard tended to spend the least on average per transaction among these three.
- Since p-value (0.0475) > 0.05, this means that we don't have enough statistical evidence to say that the average revenue per transaction is significantly different between users of Amex, Mastercard, and Visa.

Business Recommendation for Objective 6.2:

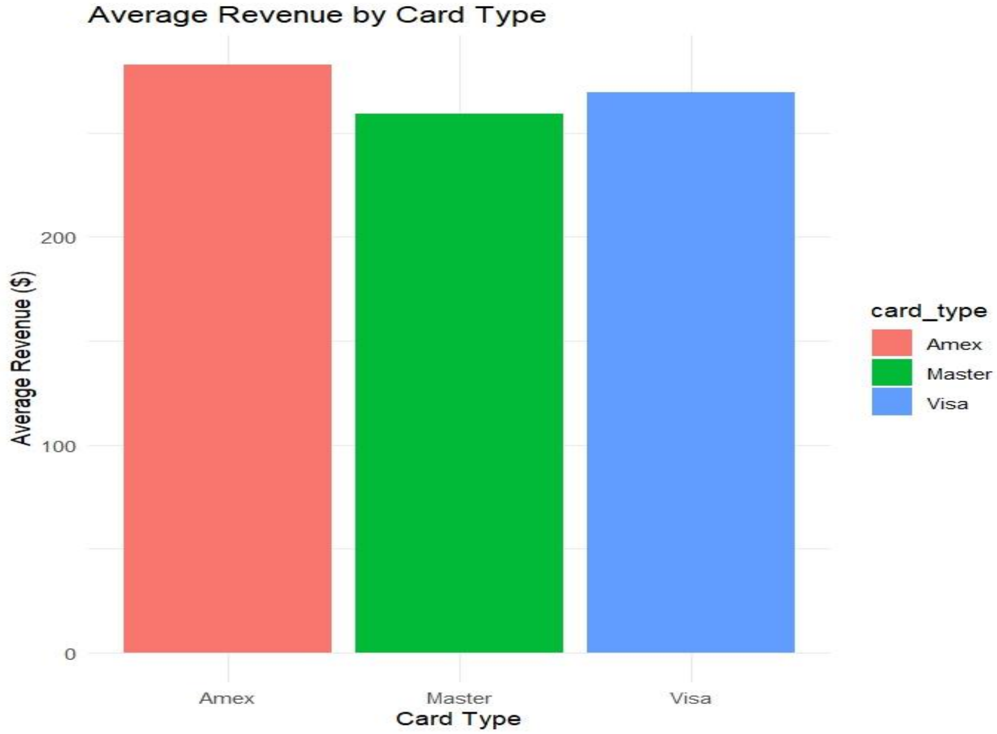
- ❑ **Accept All Major Card Types since the revenue and average spending are not significantly different, this will help to cater to a wider range of customers.**

3) Visualization: Average Revenue by Payment type



Interpretation: The bar plot shows that the highest average revenue is generated through Card payments then followed by COD.

4) **Visualization: Average Revenue by Card type**



Interpretation: This bar plot shows that Amex card generates highest average revenue among the different card types while Master generates the lowest.

6. Final Conclusions

This section includes the key takeaways from the business recommendations for each objective.

Objective 1: Optimize Landing Page Variants

- Prioritize the **Cold variant** for maximizing conversions.
- Improve **Heat and Vibrant** variants to enhance performance.

Objective 2: Optimize Demographic & Location-Based Strategies

- **Age & Gender** have minimal impact on conversions—focus on user behaviour instead.
- Allocate more resources to **high-converting cities** and investigate barriers in underperforming ones.
- Optimize **marketing strategies** in mid-tier locations to push conversions higher.
- Device type does not significantly impact conversions, but **mobile experience improvements** could be explored.

Objective 3: Optimize Traffic Sources

- Website experience optimization should be the primary focus across all traffic sources.
- Conduct **qualitative analysis** within each channel for targeted improvements.

Objective 4: Investigate Engagement Metrics

- Time spent and page visits alone aren't strong predictors—conduct **user interviews and usability testing** to uncover deeper insights.

Objective 5: Improve Coupon Strategy

- Coupons likely influence conversions, but tracking needs improvement.
- Implement **A/B testing** to test different coupon strategies.
- Gather detailed data on **coupon users vs. non-users** to refine strategies.

Objective 6: Revenue Optimization by Payment Type

- **Encourage card payments** due to higher revenue contributions.
- Optimize **COD process** as it still contributes significantly.
- Investigate **NPT(Not Provided Type) issues** to recover potential revenue.
- Accept **all major card types** to cater to a broader customer base.

7. Bibliography

- **Book:** Georgiev, G. Z. (2019). *Statistical methods in online A/B testing*. Self-Published.
- **Dataset**: https://www.kaggle.com/datasets/sandeep1080/bassburst
- **Websites:**
  - https://www.oracle.com/in/cx/marketing/what-is-ab-testing/#:~:text=A%2FB%20testing%E2%80%94also%20called,based%20on%20your%20key%20metrics.
  - https://business.adobe.com/blog/basics/learn-about-a-b-testing

8. Appendix

The codes for the project are attached in this section.

```
library(tidyverse)
library(readr)

# Load the dataset
ab_data <- read.csv(file.choose(),header=TRUE)
# Quick check for missing values in key columns
colSums(is.na(ab_data))

#Objective 1
# Summary of conversion rates by variant_group
variant_summary <- ab_data %>%
  group_by(variant_group) %>%
  summarise(
    total_users = n(),
    conversions = sum(conversion_flag),
    conversion_rate = mean(conversion_flag)
  )
print(variant_summary)

# Chi-Square test across all variants
contingency_table <- table(ab_data$variant_group, ab_data$conversion_flag)
chi_test <- chisq.test(contingency_table)
print(chi_test)

# Visualization
ggplot(variant_summary, aes(x = variant_group, y = conversion_rate, fill = variant_group)) +
  geom_bar(stat = "identity") +
  labs(title = "Conversion Rate by Variant Group", x = "Variant Group", y = "Conversion Rate") +
  theme_minimal()

#Objective 2
# Conversion rates by demographic_age_group
age_summary <- ab_data %>%
  group_by(demographic_age_group) %>%
  summarise(conversion_rate = mean(conversion_flag))
print(age_summary)

# Chi-Square test for age_group
age_table <- table(ab_data$demographic_age_group, ab_data$conversion_flag)
chi_test_age <- chisq.test(age_table)
print(chi_test_age)

# Conversion rates by demographic_gender
gender_summary <- ab_data %>%
  group_by(demographic_gender) %>%
  summarise(conversion_rate = mean(conversion_flag))
print(gender_summary)

# Chi-Square test for gender
gender_table <- table(ab_data$demographic_gender, ab_data$conversion_flag)
chi_test_gender <- chisq.test(gender_table)
print(chi_test_gender)

# Conversion rates by location (top 10 for simplicity)
location_summary <- ab_data %>%
  group_by(location) %>%
  summarise(conversion_rate = mean(conversion_flag)) %>%
  arrange(desc(conversion_rate)) %>%
  head(10)
```

```r
print(location_summary)

# Conversion rates by device_type
device_summary <- ab_data %>%
  group_by(device_type) %>%
  summarise(conversion_rate = mean(conversion_flag))
print(device_summary)

# Chi-Square test for device_type
device_table <- table(ab_data$device_type, ab_data$conversion_flag)
chi_test_device <- chisq.test(device_table)
print(chi_test_device)

# Visualization (example for age_group)
ggplot(age_summary, aes(x = demographic_age_group, y = conversion_rate, fill = demographic_age_group)) +
  geom_bar(stat = "identity") +
  labs(title = "Conversion Rate by Age Group", x = "Age Group", y = "Conversion Rate") +
  theme_minimal()

# Visualization (example for gender)
ggplot(gender_summary, aes(x = demographic_gender, y = conversion_rate, fill = demographic_gender)) +
  geom_bar(stat = "identity") +
  labs(title = "Conversion Rate by Gender", x = "Gender", y = "Conversion Rate") +
  theme_minimal()

# Visualization (example for location)
ggplot(location_summary, aes(x = reorder(location, conversion_rate), y = conversion_rate, fill = conversion_rate)) +
  geom_bar(stat = "identity") +
  coord_flip() +  # Flips the bar chart to horizontal
  labs(title = "Conversions by Location",
       x = "Location",
       y = "Number of Conversions") +
  theme_minimal() +
  scale_fill_gradient(low = "lightblue", high = "darkblue")  # Color gradient

#Visualization (for example device_type)
ggplot(device_summary, aes(x = device_type, y = conversion_rate, fill = device_type)) +
  geom_bar(stat = "identity") +
  labs(title = "Conversion Rate by Device Type", x = "Device Type", y = "Conversion Rate") +
  theme_minimal()

#Objective 3
# Summary by traffic_source
traffic_summary <- ab_data %>%
  group_by(traffic_source) %>%
  summarise(
    conversion_rate = mean(conversion_flag),
    avg_time_spent = mean(time_spent)
  )
print(traffic_summary)

# Chi-Square test for conversion_flag
traffic_table <- table(ab_data$traffic_source, ab_data$conversion_flag)
chi_test_traffic <- chisq.test(traffic_table)
print(chi_test_traffic)

# ANOVA for time_spent (continuous outcome)
anova_traffic <- aov(time_spent ~ traffic_source, data = ab_data)
summary(anova_traffic)
```

```
# Visualization
ggplot(traffic_summary, aes(x = traffic_source, y = conversion_rate, fill = traffic_source)) +
  geom_bar(stat = "identity") +
  labs(title = "Conversion Rate by Traffic Source", x = "Traffic Source", y = "Conversion Rate") +
  theme_minimal()

ggplot(traffic_summary, aes(x = traffic_source, y = avg_time_spent, fill = traffic_source)) +
  geom_bar(stat = "identity") +
  labs(title = "Average Time Spent by Traffic Source", x = "Traffic Source", y = "Time Spent (minutes)")
+
  theme_minimal()

#Objective 4
# Logistic regression for conversion_flag
logit_model <- glm(conversion_flag ~ time_spent + pages_visited, data = ab_data, family =
"binomial")
summary(logit_model)

# Visualization: Converters behaviour in terms of time spent and pages visited
ggplot(ab_data, aes(x = time_spent, y = pages_visited, color = as.factor(conversion_flag))) +
  geom_point(alpha = 0.6) +  # Scatter plot points
  geom_smooth(method = "lm", se = FALSE) +  # Linear regression trendline
  labs(title = "Time Spent vs Pages Visited (Colored by Conversion)",
       x = "Time Spent (minutes)",
       y = "Pages Visited",
       color = "Conversion") +
  theme_minimal()

# Visualization: BOxplot to compare converters vs. non-converters
# Boxplot for time spent by conversion
ggplot(ab_data, aes(x = as.factor(conversion_flag), y = time_spent, fill = as.factor(conversion_flag))) +
  geom_boxplot() +
  labs(title = "Time Spent Distribution by Conversion Status",
       x = "Conversion (0 = No, 1 = Yes)",
       y = "Time Spent (minutes)") +
  theme_minimal()

# Boxplot for pages visited by conversion
ggplot(ab_data, aes(x = as.factor(conversion_flag), y = pages_visited, fill = as.factor(conversion_flag))) +
  geom_boxplot() +
  labs(title = "Pages Visited Distribution by Conversion Status",
       x = "Conversion (0 = No, 1 = Yes)",
       y = "Pages Visited") +
  theme_minimal()

#Objective 5
# Compute Conversion Rate by Coupon Usage
coupon_analysis <- ab_data %>%
  group_by(coupon_applied) %>%
  summarise(
    conversions = sum(conversion_flag),
    total_users = n(),
    conversion_rate = conversions / total_users
  )
print(coupon_analysis)

#Chi-square test for coupons
coupon_table <- table(ab_data$coupon_applied, ab_data$conversion_flag)
chi_test_coupon <- chisq.test(coupon_table)
print(chi_test_coupon)

#Logistic regression: Coupon effect on purchase probability
```

```
logit_coupon <- glm(conversion_flag ~ coupon_applied,
                    data = ab_data,
                    family = binomial)
summary(logit_coupon)

#Compare revenue between coupon users and non-users
revenue_comparison <- ab_data %>%
  group_by(coupon_applied) %>%
  summarise(
    total_revenue = sum(revenue_.),
    avg_revenue_per_user = mean(revenue_.)
  )
print(revenue_comparison)

# Plot Coupon vs. Conversion Rate
ggplot(coupon_analysis, aes(x = as.factor(coupon_applied), y = conversion_rate, fill = as.factor(coupon_applied))) +
  geom_bar(stat = "identity") +
  labs(title = "Impact of Coupons on Conversion Rate", x = "Coupon Applied", y = "Conversion Rate") +
  theme_minimal()

#Objective 6
# Summary by payment_type
payment_summary <- ab_data %>%
  group_by(payment_type) %>%
  summarise(
    total_revenue = sum(revenue_.),
    avg_revenue = mean(revenue_.)
  )
print(payment_summary)
```

```
# ANOVA for payment_type
anova_payment <- aov(revenue_. ~ payment_type, data = ab_data)
summary(anova_payment)

# Summary by card_type (filtering for Card payments)
card_summary <- ab_data %>%
  filter(payment_type == "Card") %>%
  group_by(card_type) %>%
  summarise(
    total_revenue = sum(revenue_.),
    avg_revenue = mean(revenue_.)
  )
print(card_summary)

# ANOVA for card_type
anova_card <- aov(revenue_. ~ card_type, data = filter(ab_data, payment_type == "Card"))
summary(anova_card)

# Visualization
ggplot(payment_summary, aes(x = payment_type, y = avg_revenue, fill = payment_type)) +
  geom_bar(stat = "identity") +
  labs(title = "Average Revenue by Payment Type", x = "Payment Type", y = "Average Revenue ($)") +
  theme_minimal()

ggplot(card_summary, aes(x = card_type, y = avg_revenue, fill = card_type)) +
  geom_bar(stat = "identity") +
  labs(title = "Average Revenue by Card Type", x = "Card Type", y = "Average Revenue ($)") +
  theme_minimal()
```