Assignment 3

Analysis of “Homely Goods” data, to compare activity and expenditure between registered and non-registered users, to check whether 5% credit scheme to registered users is worth continuing.

Table of Contents

Introduction

In today’s fast-paced digital landscape, businesses are adopting measures to stay abreast with their competitors. Over the past decade, there has been a huge increase in online retail expenditure by consumers. This in turn makes business try and attract more consumers with discounts and promotional offers.

One such business “Homely Goods”, to enhance customer retention and experience, brought out a promotional scheme, of providing 5% credit on first purchases to their registered users. The purpose of this report is to research and analyse whether there is any point of continuing this promotional scheme.

Methods

To carry out this research and analysis, SAS Studio was used. Functions in SAS like PROC SQL were used to join tables, Proc TTest (with alpha at 0.02) to carry out hypothesis testing between non-registered and registered user’s activity time and spends, and plots and prints were used to visualise the data.

The following datasets were used to undertake the research and analysis:

- access_log.sas7bdat –1000 observations and 5 variables (Session_ID, Date, Start_Time, End_Time and Reg_User)
- exchange.sas7bdat –93 observations and 3 variables (date, currency, rate)
- Items.csv – 45 observations and 2 variables (Item_Code and Unit_Price)
- Order.csv – 498 observations and 4 variables (Order_Number, Session_ID, Item_Code and Quantity)
- registration.csv – 1000 observations and 7 variables (ID, First_name, Last_name, Gender, DOB, Address, Country)
- transaction.sas7bdat – 250 observations and 6 variables (Transaction, Trans_Date, Order_Number, Method, currency, Value)

Results

To carry out the analysis, first all the data was loaded onto SAS Studio and relevant library references were created to access the data easily. Once all the six datasets (Access_Log, Exchange, Items, Orders, Registration and Transaction), we checked what the variables represented.

The first thing that was evident was the table access_log failed to state whether a user was a registered or non-registered user, and what each user’s browsing time was. Using PROC SQL, two new variables were created in access_log – Status, to indicate whether a user was “Registered” or “Not Registered” and Active, to indicate the browsing time per user.

The new variable Active was quite insightful, as it showed exactly how long each user had access to the website of “Homely Goods”. A T Test was carried out to check whether the mean browsing time between the two groups (Registered/Not Registered) was the same or not. The results of the T Test are provided below.

$$H_0: Mean\ Browsing\ time\ of\ Registered\ users - mean\ browsing\ time\ of\ non\ registered\ users = 0$$

Two-sample t-test to compare browsing time for the two versions

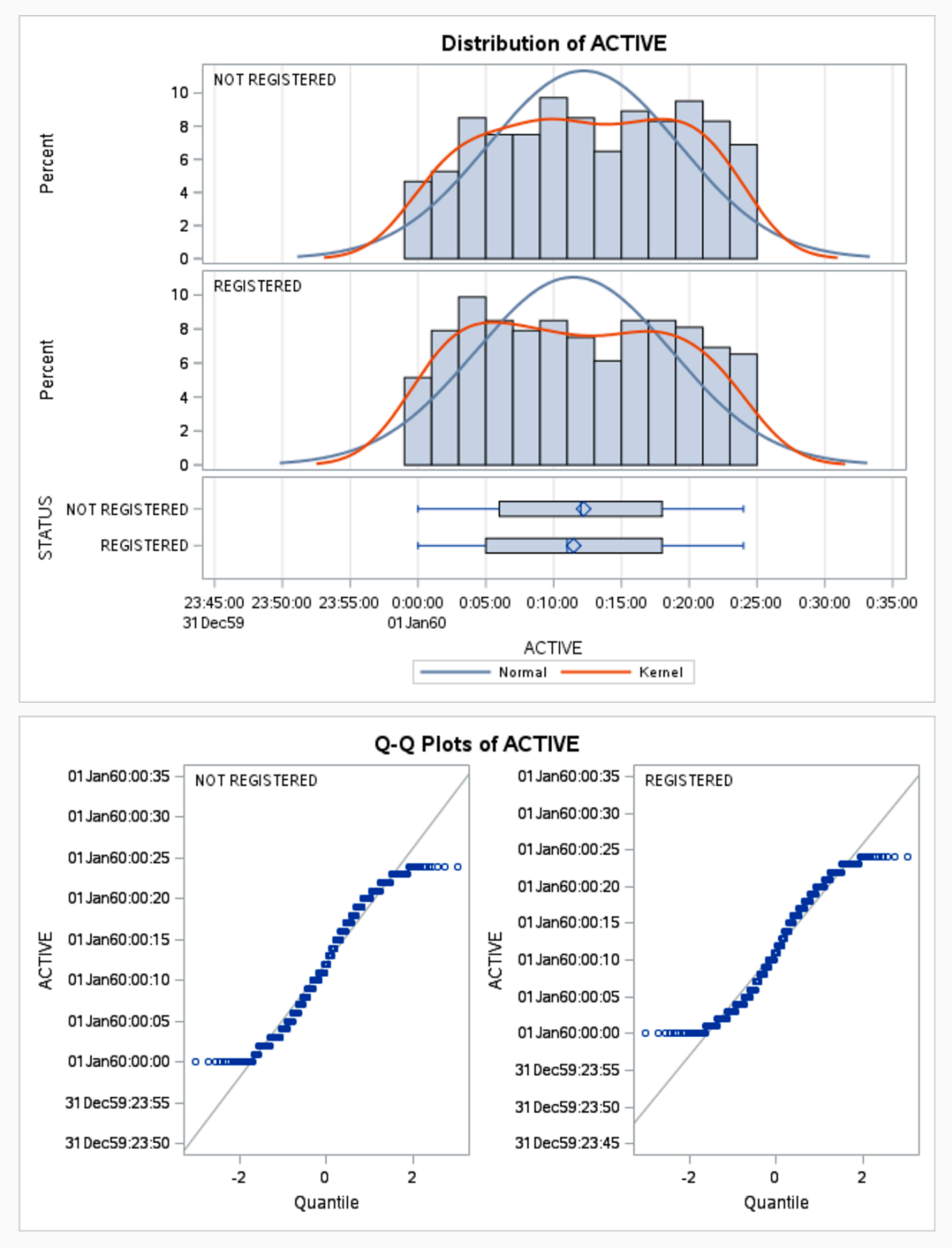
The TTEST Procedure
Variable: ACTIVE

STATUS	N	Mean	Std Dev	Std Err	Minimum	Maximum
NOT REGISTERED	494	732.6	422.0	18.9884	0	1440.0
REGISTERED	506	688.3	433.2	19.2578	0	1440.0
Diff (1-2)		44.2917	427.7	27.0533		

STATUS	Method	Mean	98% CL Mean		Std Dev	98% CL Std Dev	
NOT REGISTERED		732.6	688.3	776.9	422.0	392.9	455.7
REGISTERED		688.3	643.4	733.3	433.2	403.6	467.3
Diff (1-2)	Pooled	44.2917	-18.7450	107.3	427.7	406.5	451.2
Diff (1-2)	Satterthwaite	44.2917	-18.7253	107.3			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	998	1.64	0.1019
Satterthwaite	Unequal	998	1.64	0.1018

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	505	493	1.05	0.5606



From the T-Test we can infer that since the p value is greater than the significance level of alpha = 0.02, hence we fail to reject the null hypothesis that the average browsing time is equal. There is thus no statistically significant result to support this query.

Next, in the transaction table it was evident that there were transactions in multiple currencies, and comparing spends would be impossible. To make the comparison between expenditures of the two user groups, tables exchange and transaction were joined, to correctly convert all values into Australian Dollars (AUD) based on rates by the date of transaction.

Now that all values of transactions were in AUD, we further carried out another T-Test to check whether the mean expenditure between the two user groups were equal. For this hypothesis test, we again use alpha = 0.02.

$$H_0: Mean\ Expenditure\ of\ Registered\ users - Mean\ Expenditure\ of\ non\ registered\ users = 0$$

The results of the T-Test are below.

Two sample t-test to compare money received for the two groups. (USING α=0.02)

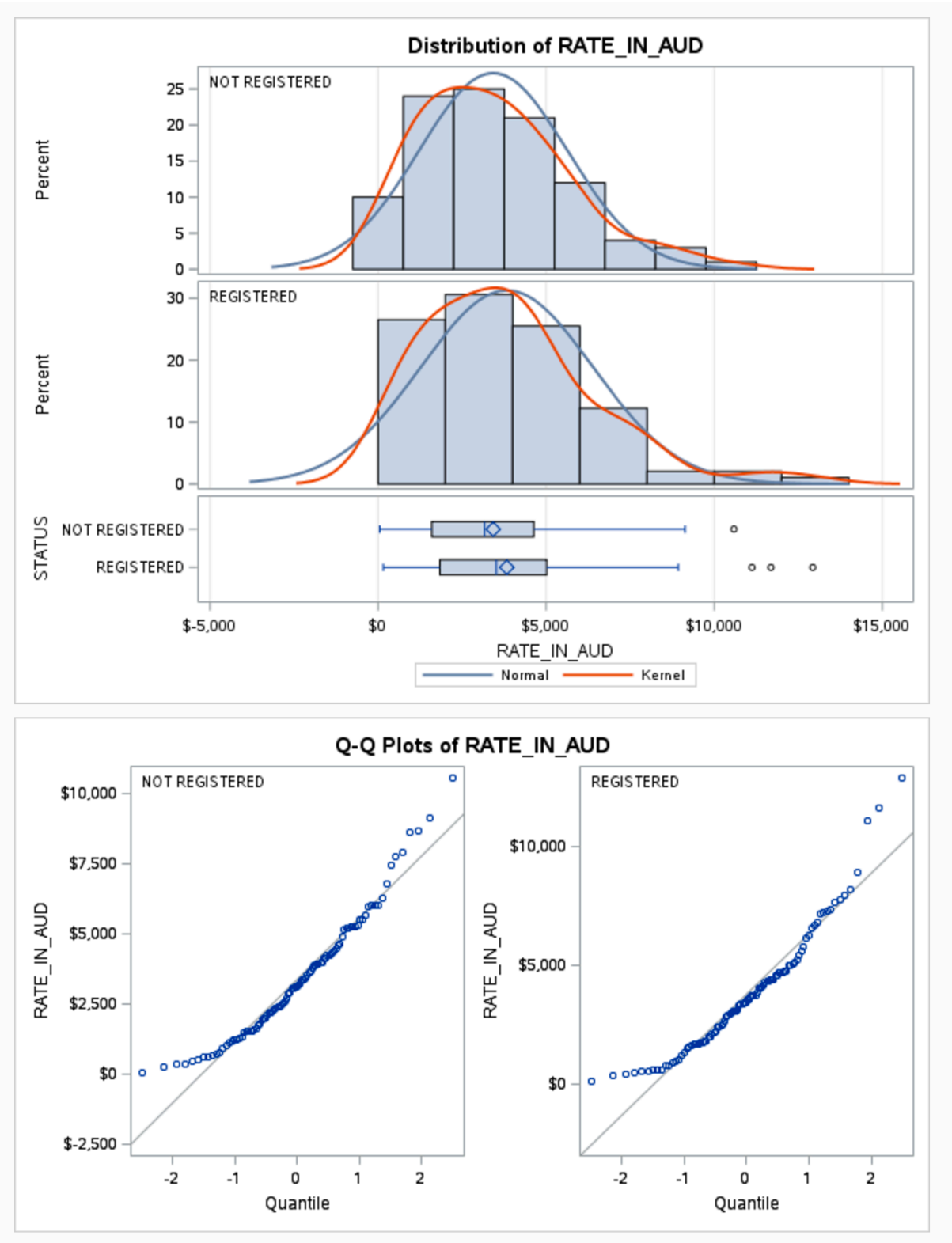
The TTEST Procedure
Variable: RATE_IN_AUD

STATUS	N	Mean	Std Dev	Std Err	Minimum	Maximum
NOT REGISTERED	100	3422.4	2199.2	219.9	47.9685	10585.5
REGISTERED	98	3829.1	2551.5	257.7	157.4	12927.4
Diff (1-2)		-406.7	2380.1	338.3		

STATUS	Method	Mean	98% CL Mean		Std Dev	98% CL Std Dev	
NOT REGISTERED		3422.4	2902.4	3942.4	2199.2	1885.8	2629.9
REGISTERED		3829.1	3219.5	4438.8	2551.5	2184.7	3057.2
Diff (1-2)	Pooled	-406.7	-1200.2	386.8	2380.1	2128.9	2694.7
Diff (1-2)	Satterthwaite	-406.7	-1201.6	388.2			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	196	-1.20	0.2307
Satterthwaite	Unequal	190.65	-1.20	0.2315

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	97	99	1.35	0.1427



From the T-Test we can infer that since the p value is greater than the significance level of alpha = 0.02, hence we fail to reject the null hypothesis that the mean expenditure is equal. There is thus no statistically significant result to support this query.

Next, to find out how many registered users received credit we had to filter our access log based on registered users without any duplicate key, to ensure the first activity of the user gets logged. Then filtering of the table was carried out to ensure only registered users were given a 5% credit based on their first transaction since 1st July.

From the results of this, we could see that 94 unique registered users received 5% credit for their purchases.

To understand whether this credit scheme favoured the business, a comparative table between the total browsing time and total expenses between the two user groups were created and have been added below.

Table and pie chart showcasing the total expenditure by the two user groups.

Obs	STATUS	SPENDS
1	NOT REGISTERED	$342,241.99
2	REGISTERED	$375,256.39

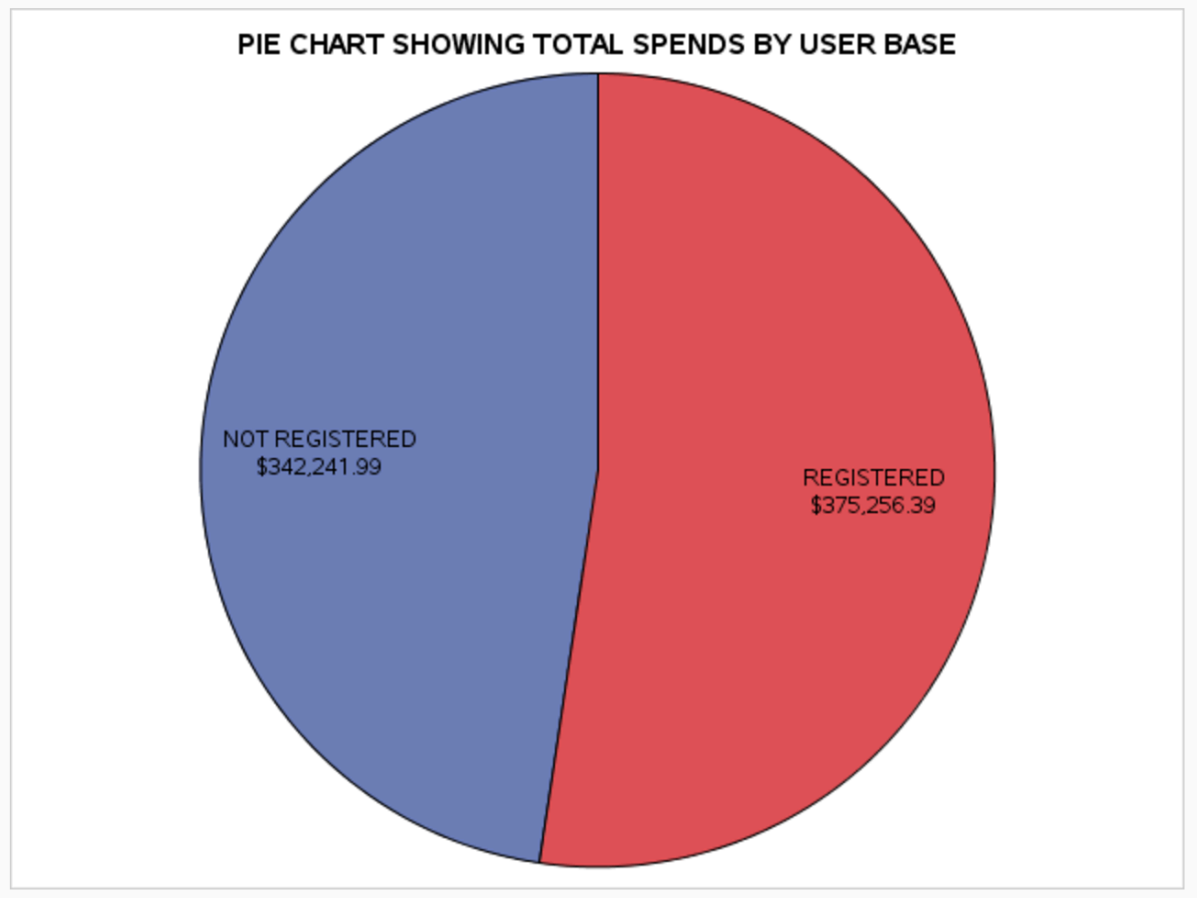


Table showcasing total and average browsing time between the two user groups.

Obs	STATUS	TOTAL_BROWSING_TIME	AVERAGE_BROWSING_TIME
1	NOT REGISTERED	100:32:00	0:12:13
2	REGISTERED	96:45:00	0:11:28

From the above tables and pie chart the following points were inferred:

1. Registered users spent more than non registered users by approximately AUD 33,000
2. Non registered users spent more time than registered users on the website.

Discussion

Although the results of the hypothesis tests point towards there being no statistically significant results in the differences in spends and browsing time between groups, the summaries and tables show that registered users indeed spent more.

There were many limitations to this research as follows:

1. In the transaction table, out of 250, 52 transactions were not linked with the order table, hence it is unknown as to whether they were purchased.
2. The data provided is for one month, and hence it is not clear whether the credit scheme attracted new users.

Conclusion

The results of the T-Tests were not statistically significant in comparing the mean browsing times and mean expenses between two user’s groups, and we failed to reject the null hypothesis.

Since 94 registered users could obtain credit from their purchases, it is a good scheme. In fact, if “Homely Goods” can continue this scheme to convert non-registered users, it would add to their customer base.

References

- Data sourced from “Homely Goods” and modified for the purpose of this assignment.
- https://www.forbes.com/sites/daniellemarceau/2017/01/09/where-is-all-the-money-going-a-look-at-consumer-spending-in-2017/#250af44625cf

Appendix

2017-5-31 Code: ASSIGNMENT3 - S3623575.sas

```
/* Q1. SETTING LIBREF AND READING IN CSV FILES TO SAS */
LIBNAME ASS3 "/home/s36235750/sasuser.v94/Assignment 3";

LIBNAME ASS3OUT "/home/s36235750/sasuser.v94/Assignment 3/OUTPUT";

DATA ASS3.ITEMS;
    INFORMAT ITEM_CODE $11. UNIT_PRICE COMMA6.2;
    INFILE "/home/s36235750/sasuser.v94/Assignment 3/Items.csv" DELIMITER = "," DSD MISSOVER FIRSTOBS =2;
    INPUT ITEM_CODE UNIT_PRICE;
    FORMAT ITEM_CODE $11. UNIT_PRICE COMMA6.2;
RUN;

DATA ASS3.ORDER;
    INFORMAT ORDER_NUMBER 3. SESSION_ID 4. ITEM_CODE $11. QUANTITY 2.;
    INFILE "/home/s36235750/sasuser.v94/Assignment 3/Order.csv" DELIMITER = "," DSD MISSOVER FIRSTOBS =2;
    INPUT ORDER_NUMBER SESSION_ID ITEM_CODE QUANTITY;
    FORMAT ORDER_NUMBER $3. SESSION_ID $3. ITEM_CODE $11. QUANTITY $2.;
RUN;

DATA ASS3.REGISTRATION;
    INFORMAT ID 4. FIRST_NAME $16.  LAST_NAME $16. GENDER $6. DOB DDMMYY10. ADDRESS $45. COUNTRY $25.;
    INFILE "/home/s36235750/sasuser.v94/Assignment 3/registration.csv" DELIMITER = "," DSD MISSOVER FIRSTOBS =2
    INPUT ID FIRST_NAME LAST_NAME GENDER DOB ADDRESS COUNTRY;
    FORMAT ID 4. FIRST_NAME $16.  LAST_NAME $16. GENDER $6. DOB DDMMYY10. ADDRESS $45. COUNTRY $25.;
RUN;


/* Q2. CHECKING ACCESS_LOG AND CREATING STATUS VARIABLE TO REPRESENT REGISTER AND NON-REGISTERED USERS */
PROC CONTENTS DATA = ASS3.ACCESS_LOG ORDER = VARNUM;
RUN;

PROC SQL;
CREATE TABLE ASS3.ACCESS_LOGV2 AS
SELECT *,
(CASE
when REG_USER IS NOT NULL then "REGISTERED"
when REG_USER IS NULL then "NOT REGISTERED" else ""
end) AS STATUS
FROM ASS3.ACCESS_LOG;
QUIT;


/* Q3. Create a new variable “Active” to represent the browsing time for users on the website (in hours
and minutes, to the nearest minute using an appropriate time format) for each visit. */

PROC SQL;
CREATE TABLE ASS3.ACCESS_LOGV3 AS
SELECT *, (END_TIME - START_TIME) FORMAT TIME5. AS ACTIVE
FROM ASS3.ACCESS_LOGV2;
QUIT;


/* Q4. Two-sample t-test to compare browsing time for the two versions. Use α=0.02. */
PROC TTEST DATA = ASS3.ACCESS_LOGV3 ALPHA = 0.02 ;
TITLE 'Two-sample t-test to compare browsing time for the two versions (USING α=0.02)';
CLASS STATUS;
VAR ACTIVE;
RUN;


/* Q5. Calculate the amount of money that has been received for each order in AUD. */

PROC SQL;
CREATE TABLE ASS3.TRANSACTIONV2 AS
SELECT T.TRANSACTION, T.TRANS_DATE, T.ORDER_NUMBER, T.METHOD, T.CURRENCY, T.VALUE,
(CASE
WHEN T.CURRENCY = "CAD" THEN T.VALUE / E.RATE
WHEN T.CURRENCY = "NZD" THEN T.VALUE / E.RATE
WHEN T.CURRENCY = "GBP" THEN T.VALUE / E.RATE
ELSE T.VALUE*1
END) FORMAT DOLLAR10.2 AS RATE_IN_AUD
FROM ASS3.TRANSACTION T
LEFT JOIN ASS3.EXCHANGE E
ON T.CURRENCY = E.CURRENCY
AND T.TRANS_DATE = E.DATE;
QUIT;


/* Q6. Run a two sample t-test to compare money received for the two groups. Use α=0.02. */

PROC SQL;
CREATE TABLE ASS3.ACCESS_LOGV4 AS
SELECT DISTINCT TV2.ORDER_NUMBER, O.SESSION_ID, TV2.TRANSACTION, TV2.TRANS_DATE,
TV2.RATE_IN_AUD, AL3.REG_USER, AL3.STATUS
FROM ASS3.TRANSACTIONV2 TV2
LEFT JOIN ASS3.ORDER O
```

1/3

```
ON O.ORDER_NUMBER = TV2.ORDER_NUMBER
LEFT JOIN ASS3.ACCESS_LOGV3 AL3
ON AL3.SESSION_ID = O.SESSION_ID
HAVING O.ORDER_NUMBER = TV2.ORDER_NUMBER;
QUIT;


PROC TTEST DATA = ASS3.ACCESS_LOGV4 ALPHA = 0.02;
TITLE 'Two sample t-test to compare money received for the two groups. (USING α=0.02)';
CLASS STATUS;
VAR RATE_IN_AUD;
RUN;


/* Q7. Calculate the possible credit that may be redeemed by the registered users assuming the
scheme started on 1 July. */

PROC SORT DATA = ASS3.ACCESS_LOGV4 NODUPKEY;
BY REG_USER;
RUN;

PROC SQL;
CREATE TABLE ASS3.ACCESS_LOGV5 AS
SELECT REG_USER, TRANS_DATE, RATE_IN_AUD, (RATE_IN_AUD * .05) FORMAT DOLLAR10.2 AS CREDIT
FROM ASS3.ACCESS_LOGV4
WHERE STATUS = "REGISTERED"
;
QUIT;

PROC SQL;
CREATE TABLE ASS3.REGISTRATIONV2 AS
SELECT R.ID, R.FIRST_NAME, R.LAST_NAME, R.GENDER, R.DOB, R.ADDRESS, R.COUNTRY, ALV5.CREDIT
FROM ASS3.REGISTRATION R
LEFT JOIN ASS3.ACCESS_LOGV5 ALV5
ON R.ID = ALV5.REG_USER
;
QUIT;


/* Q8. any other relevant analysis of this data source */

/* CREATING A NEW TABLE TO COMBINE STATUS VARIABLE IN TRANSACTION TABLE */
PROC SQL;
CREATE TABLE ASS3.TRANSACTIONV3 AS
SELECT UNIQUE TV2.TRANS_DATE, TV2.TRANSACTION, TV2.ORDER_NUMBER, ALV2.STATUS, TV2.RATE_IN_AUD
FROM ASS3.TRANSACTIONV2 TV2
LEFT JOIN ASS3.ORDER O
ON TV2.ORDER_NUMBER = O.ORDER_NUMBER
LEFT JOIN ASS3.ACCESS_LOGV2 ALV2
ON O.SESSION_ID = ALV2.SESSION_ID
HAVING TV2.ORDER_NUMBER = O.ORDER_NUMBER
;
QUIT;

/* COMPARING TOTAL SPENDS BY REGISTERED USERS AND NON REGISTERED USERS */
PROC SQL;
CREATE TABLE ASS3.SALESBYUSERS AS
SELECT TV3.STATUS, SUM(TV3.RATE_IN_AUD) FORMAT DOLLAR11.2 AS SPENDS
FROM ASS3.TRANSACTIONV3 TV3
GROUP BY STATUS;
QUIT;

/* CREATING A TABLE WITH TOTAL AND AVERAGE BROWSING TIMES BY USERS BASE */
PROC SQL;
CREATE TABLE ASS3.BROWSINGTIMEBYUSERS AS
SELECT AL3.STATUS, SUM(AL3.ACTIVE) FORMAT TIME9. AS TOTAL_BROWSING_TIME,
MEAN(AL3.ACTIVE) FORMAT TIME9. AS AVERAGE_BROWSING_TIME
FROM ASS3.ACCESS_LOGV3 AL3
GROUP BY STATUS;
QUIT;

/* CREATING A VISUALISATION TO SHOW DIFFERENCE IN SPENDS BY USER BASE */
proc template ;
    define statgraph WebOne.Pie;
        begingraph;
        entrytitle "PIE CHART SHOWING TOTAL SPENDS BY USER BASE";
        layout region;
        piechart category=STATUS response=SPENDS / start=90;
        endlayout;
        endgraph;
    end;
run;

/*--Set output size--*/
ods graphics / reset imagemap;

/*--SGRENDER proc statement--*/
```

2017-5-31 Code: ASSIGNMENT3 - S3623575.sas

```
proc sgrender template=WebOne.Pie data=ASS3.SALESBYUSERS;
run;

ods graphics / reset;
```

3/3