Q1	No. of observations Provided correct info	No. of observations Not Provided	No. of observations Provided incorrect info	Total No. of observations
	3930 (64.5%)	890 (14.6%)	1270 (20.9%)	6090 (100%)
	3930 (64.5%)	2160 (35.5%)		6090 (100%)
	3930 (75.6%)	Removed	1270 (24.4%)	5200 (100%)
	The site investigated 203 members with 30 items, total observation is 6090, after remove “Not Provided”, observation is 5200			

Var	Description	Total Observation n_1 :6090	After remove n_2 :5200	Findings
Knowledge of the site	Yes	3150 (51.7%)	2725 (52%)	More than half observations (52%) of user has knowledge on the site.
	No	2940 (48.3%)	2475 (48%)	
Experience	Yes	2520 (41.4%)	2169 (42%)	More than half observations (58%) of user has direct or indirect experience of online matchmaking. It means users have already built a trust relationship on a site.
	No	3570 (58.6%)	3031 (58%)	
Internet trust	1 (Don’t believe)	30 (0.5%)	30 (0.6%)	Around 37% observations of user have got moderate trustworthy of the internet.
	>1-2	240 (3.9%)	211 (4.1%)	
	>2-3	1560 (26%)	1390 (27%)	
	>3-4	2370 (39%)	1921(37%)	
	>4-5	1680 (28%)	1447 (28%)	
	>5-6 (High)	210 (3.4%)	201(3.9%)	
Internet privacy concern	>1-2 (Low)	90(1.5%)	88 (1.7%)	Around 82% observations of user have got moderate to high concern about Internet privacy.
	>2-3	180 (3%)	*154* (3%)	
	>3-4	840 (14%)	708 (14%)	
	>4-5	2010 (33%)	1761(34%)	
	>5-6	2010 (33%)	1657 (32%)	
	>6-7(High)	960 (17%)	832 (16%)	
Age	18	30 (0.5%)	23(0.4%)	Around 82% observations of user age distribution are between 19-24.
	19-20	1470 (24%)	1200 (23%)	
	21-22	1680 (28%)	1432 (28%)	
	23-24	1830 (30%)	1562 (30%)	
	25-26	810 (13%)	733 (14%)	
	27-28	150 (2.5%)	133 (2.6%)	
	29-30	120 (2%)	117 (2.3%)	
sex	Female	2820 (46%)	2381(46%)	More than half observations (54%) of user is male.
	male	3270 (54%)	2819 (54%)	
Provide correct info	1 (correct)	3930 (64.5%)	3930 (75.6%)	More than 75% observations of user have provided correct information.
	2 (not provided)	890 (14.9%)	-	
	3 (incorrect)	1270 (20.9%)	1270 (24.4%)	
Sensitivity	0 (Not at all)	0	0	Around 21% observations of user believed that the items of survey are moderately sensitive to provide information.
	1-10	0	0	
	11-20	0	0	
	21-30	812 (13%)	719 (14%)	
	31-40	1015 (17%)	917 (18%)	
	41-50	1218 (20%)	1095 (21%)	
	51-60	406 (6.7%)	352 (6.8%)	
	61-70	812 (13%)	619 (12%)	
	71-80	1218 (20%)	980 (19%)	
	81=90	406 (6.7%)	344 (6.6%)	
	91-100 (extremely)	203 (3.3%)	174 (3.3%)	
Relevant	0 (Not at all)	0	0	More than 61% observations of user believed that the items of survey are moderately to high relevant (51-90) to find a good matching.
	1-10	0	0	
	11-20	203 (3.3%)	174 (3.3%)	
	21-30	406 (6.7%)	365 (7%)	
	31-40	812 (13%)	678 (13%)	
	41-50	1015 (17%)	787 (15%)	
	51-60	812 (13%)	752 (14%)	
	61-70	1421 (23%)	1151(22%)	
	71-80	1015 (17%)	916 (18%)	
	81=90	406 (6.7%)	377 (7.3%)	
	91-100 (extremely)	0	0	

(Cont'Q1) Hypothesis	Observations	t-statistic	p-value	Findings by t-test
H1: There are sensitivity difference between those provided correct info with those not provided & provided incorrect info.	6090	- 23.328	p< 2.2e-16 (sig. & supported)	They are significantly different in terms of sensitivity. Less sensitive (mean= 49.5) in those have provided correct info than those not provided+ provided incorrect info (mean= 61.7).
Finding: The accuracy of providing information (D=1) is highly depend on contextual nature of information sensitivity in survey.	5200 (removed not provided)	-22.579	p< 2.2e-16 (sig. & supported)	They are significantly different in terms of sensitivity. Less sensitive (mean= 49.5) in those have provided correct info than those provided incorrect info (mean= 64.1).
H2: There are relevant difference between those provided correct info with those not provided +provided incorrect info.	6090	19.482	p< 2.2e-16 (sig. & supported)	They are significantly different in terms of relevant. Much relevant (mean= 58.5) in those have provided correct info than those not provided+ provided incorrect info (mean=49.4).
Finding: Relevance of information asked in survey will positively influence user attitudes toward provide correct information (D=1).	5200 (removed not provided)	21.395	p-value < 2.2e-16(sig. & supported)	They are significantly different in terms of relevant. Much relevant (mean= 58.5) in those have willing provided correct info than those provided incorrect info (mean=46.2).

After performing statistical analysis, we decide to remove "not provided" observations for this assessment, because (i) we don't have any information for "not provided observations" (ii) Even removed "not provided" observations, the dataset is still good enough (still maintain high population (n=5200) with likely no bias compared with original one.

Q2 & Q3. Pooling model ignores the unobserved heterogeneity of users, and it also ignores possible association within users (groups). To control for the unobserved heterogeneity, we can take the first difference (period-to-period change) and use it for the analysis by First Difference (FD) Model, or include the dummies indicating each user by Fixed Effects (FE) Model (or Least Squares Dummy Variable (LSDV) Model) Or use a group(user)-specific random element by Random Effects (RE) Model.

FD model perfectly controls for time-invariant user heterogeneity.
FE Model: Make the dummy variables indicating each user and apply the regression to the data including all dummy variables. It's perfectly controls for time-invariant user heterogeneity.
RE model separates individual effects into two pieces: (a) individual effects from observed heterogeneity; (b) individual effects from unobserved heterogeneity, which is a group(user)-specific random element and has a strict assumption that the effects are uncorrelated with the regressors. RE model cannot perfectly controls for time-invariant user heterogeneity

Typically, FE model is preferred to RE model, but we can test the observed heterogeneity as it can be incorporated in the RE model. Also, we can save the degree of freedom so we can have a better chance of rejecting the null hypothesis. (See Table in p.3)

Q4		F-statistic	p-value	# Findings1	Wald test
	Main effect model	622.5	< 2.2e-16(sig. & supported)	There is main effect model of relv on the relationship between D & sen.	The purpose of Wald test is to find out if explanatory variables in a model are significant. We found that Chi-squared X2 is 11.4 & P(> X2) 0.00075. It shows the parameters are not zero, which means we should include variables in the Moderation effect model.
	Moderation effect model	419.6	< 2.2e-16(sig. & supported)	There is moderation effect model of relv on the relationship between D & sen.	
	Main effect	$Relv$ = 66.35877 − 0.31457 sen +7.75973 D			
	Moderation effect		$Relv$ = 70.38979 − 0.37747 sen + 2.44947 D + 0.08874 sen*D		

To test the moderation effect, we include the interaction term between the predictor and the moderator, and test the coefficient of the interaction. $relv = c + \beta_1 sen + \beta_2 D + \beta_3 senD + \gamma_1 z_1 + \gamma_2 z_2 + \cdots + \varepsilon$
Then, we have two models for fitting the data. $relv = c + \beta_1 sen + \beta_2 + \beta_3 sen + \gamma_1 z_1 + \gamma_2 z_2 + \cdots + \varepsilon \qquad when\ D = 1$

$$relv = c + (\beta_1 + \beta_{3)} sen + \beta_2 + \gamma_1 z_1 + \gamma_2 z_2 + \cdots + \varepsilon \qquad when\ D = 1$$

$$relv = c + \beta_1 sen + \gamma_1 z_1 + \gamma_2 z_2 + \cdots + \varepsilon \qquad when\ D = 0$$

Compare the fit (R-square) between the restricted model and the unrestricted model

$$relv = c + \beta_1 sen + \beta_2 D + \beta_3 sen\ D + \gamma_1 z_1 + \gamma_2 z_2 + \cdots + \varepsilon \qquad unrestricted\ model$$

$$relv = c + \beta_1 sen + \beta_2 D + \gamma_1 z_1 + \gamma_2 z_2 + \cdots + \varepsilon \qquad restricted\ model\ by\ \beta_3 = 0$$

Further study and Suggestions for (i) screening &(ii) main effect and interaction effect on relv, sen and D by ANOVA: From results of ANOVA, both result of Pr(>F) are very close to zero, which mean "relv" is a significant predictor, while the interaction of relv, sen and D is also significant. Furthermore, we performed model comparison in AVOVA. Two models are found significant, namely (i) Model 2 which is the significant on relv, sen and D; (ii) Model 5 which is the significant on relv, sen, D, know, expr & itr. See more details in R Markdown (Assignment 2b.Rmd) with R-Studio which including R-code and visualization graphs.

Q5. We do need more data to establish the relationship between D and sen, relv. The possible control variables like citizen, born of country, race and skin colour.

$$lnD_{it} = \beta_1 + \beta_2 sen_{it} + \beta_3 relv_{it} + \beta_4 know_{it} + \beta_5 expr_{it} + \beta_6 itr_{it} + \beta_7 ip_{it} + \beta_8 sex_i + \beta_9 age_{it} + \varepsilon_{it}$$

Table Estimate the impact of accuracy for providing information (D) on sensitivity, relevant and other control variables.																
Cont' Q2	OLS							Oneway (individual) effect First-Difference Model (FD)			Fixed effect (FE) model			Oneway (individual) effect Random Effect Model (RE)		
	OLS			Panel Robust Mode		Pooling model										
Variables	Estimates Coefficient	Std. Error	p-value	Std. Err	p-value	Std. Error	p-value	Estimates Coefficient	Std. Error	p-value	Estimates Coefficient	Std. Error	p-value	Estimates Coefficient	Std. Error	p-value
β1: constant	0.4591604	0.0701880	6.66e-1 ***	0.0713	<0.0001	0.0701879	6.660e-11 ***	0.00047534	0.00700171	0.945877	-	-	-	0.4691976	0.14557733	0.00127 **
β2: sen	-0.005088	0.0003018	< 2e-16	0.0003	<0.0001	0.0003017	< 2.2e-16	-	-	-	-	-	-	-0.005547	0.00135569	4.338e-(***
β3: relv	0.0047139	0.0003309	< 2e-16	0.0003	<0.0001	0.0003308	< 2.2e-16	-	-	-	-	-	-	0.0045382	0.00150220	0.00253 **
β4: know	-0.065745	0.0185931	0.00041 ***	0.0196	0.0008	0.0185930	0.0004098 ***	-0.0313228	0.01704313	0.066140	-0.0644196	0.01721600	0.0001847	-0.064476	0.01723934	0.00018 ***
β5: expr	0.0488335	0.0188836	0.00974	0.0204	0.0166	0.0188836	0.0097361	0.03676882	0.01747344	0.035403	0.04245882	0.01748623	0.0152109	0.0427640	0.01750988	0.01462
β6: itr	0.0588246	0.0062304	< 2e-16	0.0063	<0.0001	0.0062304	< 2.2e-16	0.04102547	0.00550429	1.061e-1 ***	0.05729759	0.00576814	< 2.2e-16 *	0.0573746	0.00577600	< 2.2e-1 ***
β7: ip	-0.000616	0.0052656	0.90677	0.0055	0.9102	0.0052655	0.9067747	-0.0060394	0.00493095	0.220705	0.00064785	0.00487486	0.8942800	0.0005875	0.00488150	0.90419
β8: sex	0.0764637	0.0119982	2.01e-1(***	0.0117	<0.0001	0.0119982	2.015e-10 ***	0.03106479	0.01105206	0.004961	0.07587051	0.01111200	6.8278 9.6 12 ***	0.0759005	0.01112695	1.005e-1 ***
β9: age	0.0029821	0.0024950	0.23205	0.0025	0.2273	0.0024950	0.2320509	0.00846236	0.00230448	0.000243 ***	0.00365518	0.00231014	0.1136568	0.0036230	0.00231327	0.11736
Finding#1	Two of Variables, "know" and "expr" of VIF are > 2.8. It means they are slightly correlated with at least one of the other predictors in the model.							To control for the unobserved heterogeneity, we can take the first difference (period-to-period change) a use it for analysis.			p-value of "ip" and "age" >0.05, so they are not significant in the model & just control variables.			p-value of "ip" and "age" >0.05, so they are not significant in the model just control variables.		
Finding#2	"ip" and "age" are just control variables.							Adv: t can remove the latent heterogeneity from the model whether the fixed or random effect model is appropriate.								
Finding#3	BP test has a p-value< 2.2e-16, so we can reject the null hypothesis and infer that heteroscedasticity is indeed present.							Disadv: The differencing also remove any time-invariant variables from the model.								
Findings	#1. We found that "ip" and "age" are not significant in above models. #2.As a general observation, with a variety of approaches available, FD estimator does not have much recommend. #3. Conduct Hausman test to check whether the individual effects are uncorrelated with the regressors, we found that p-value is 0.9979, therefore we use Random Effect Model (RE).															

Using the panel glm model, Maximum Likelihood estimation

Variables	β1: const	β2: sen	β3: relv	β4: know	β5: expr	β6:itr	β7: ip	β8: sex	β9: age	sigma	Finding
Estimate Coef	-1.012507	-0.034665	0.032038	-0.374817	0.360162	0.456089	-0.016133	0.695291	0.041967	1.826031	p-value of "sen" "relv", "itr","sex" <0.05, so they are significant in the model.
Std. Error	1.247833	0.002207	0.002386	0.381473	0.380481	0.131381	0.105941	0.222851	0.050043	0.127687	
p-value	0.417129	< 2e-16 ***	< 2e-16 ***	0.325828	0.343845	0.000518 ***	0.878966	0.001809 *	0.401679	< 2e-16 ***	