Assignment 1 in CA259

Submitted to: Prof. Alan Smeaton

Submitted by: Sebastian Gaycken (22108416)

Word Count Report: 1497 = 1047 + 3 * 150

Submitted on: 18.02.2024

Data Normalisation and Cleaning

Personal Preferences

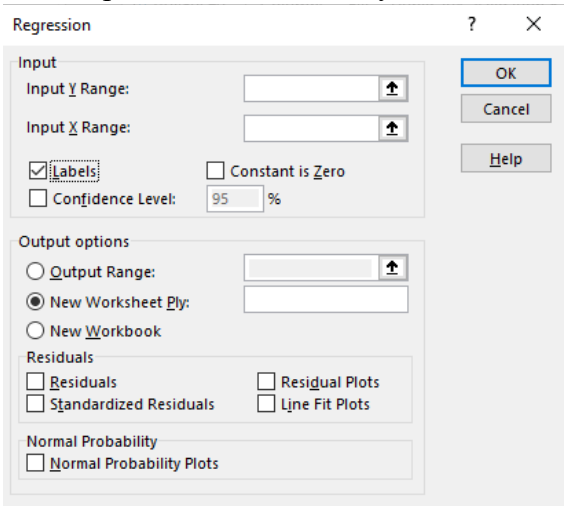
First, both sheets were converted into tables to enable easier filtering and sorting. Next, in the 'Demographics' sheet, the ID (i.e. the last four digits) was moved to the left and the format was changed to display four digits. This was necessary because Excel was removing leading zeros (e.g. 97 became 0097), which would have implied that people have phone numbers with just two digits, which is highly unlikely. Additionally, the columns were sorted into two categories: numeric (adjacent to the ID column) and non-numeric (to the right), to facilitate linear regression analysis, which only allows for numeric data. Columns with missing values were grouped by colour and placed in the last five rows. Finally, the Excel Add-in 'Data Analysis' was activated to simplify the following steps.

Imputation

Before beginning the imputation process, several considerations need to be addressed. There are five rows, each with one missing value. It is important to determine the order in which these rows will be calculated and whether the values calculated first will be used to calculate the others. If the missing values are random, then the order of calculation should not affect the outcome. The missing values in this data set were imputed in an arbitrary order (Order ID: 6290 -> 5262 -> 7677 -> 7181 -> 0838). To calculate the other values, rows with imputed values were used, which may introduce bias. However, this was deemed acceptable due to the small size of the data set and the fact that all present values in those rows would go unused.It is unclear whether non-numeric data will be used for imputation. Although theoretically, categorical data could be transformed into dummy variables for calculation, this was not discussed in class. The non-numeric data will not be used for the normalization part of this exercise.

Imputation of the age of ID 6290

1. The Data Analysis Add-in window was opened. The Input Y Range is the complete ‘Age’ column only including all complete rows. For the Input X Range, all numeric columns were included except the ID column and ‘Age’. Labels were included in those inputs for the readability of the results. All other boxes were marked like this:



2. The results of the first regression looked liked this:

	A	B	C	D	E	F	G	H	I
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	10	948.2500533	94.82500533	4.623359742	1.86701E-05			
13	Residual	106	2174.057639	20.50997773					
14	Total	116	3122.307692						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	15.99869677	10.69174353	1.496359946	0.137531482	-5.198724673	37.19611821	-5.198724673	37.19611821
18	CAO Points (100 to 600)	-0.01040857	0.005839939	-1.782307962	0.077562078	-0.021986817	0.001169677	-0.021986817	0.001169677
19	Daily travel to DCU (in km, 0 if on-campus)	0.020110755	0.028782053	0.698725533	0.486253788	-0.036952464	0.077173974	-0.036952464	0.077173974
20	Average year 1 exam result (as %)	0.061796416	0.05131943	1.204152416	0.231212645	-0.039949346	0.163542178	-0.039949346	0.163542178
21	Seat row in class	0.407349448	0.165292121	2.464421448	0.015330408	0.079641742	0.735057154	0.079641742	0.735057154
22	Number of older siblings	1.537640226	0.402501138	3.820213359	0.00022506	0.739642571	2.33563788	0.739642571	2.33563788
23	Number of younger siblings	1.4920429	0.471585646	3.163885313	0.002031741	0.557078489	2.427007312	0.557078489	2.427007312
24	Old Dublin postcode (0 if outside Dublin)	-0.057583703	0.07193891	-0.800452809	0.425238886	-0.200209589	0.085042184	-0.200209589	0.085042184
25	Height (in cm)	-0.013200969	0.073675711	-0.179176674	0.85814102	-0.159270232	0.132868294	-0.159270232	0.132868294
26	Weight (in kg)	0.046682267	0.054801269	0.851846451	0.396219778	-0.06196658	0.155331114	-0.06196658	0.155331114
27	Shoe size	0.069911619	0.267189758	0.261655311	0.794094949	-0.459818063	0.599641301	-0.459818063	0.599641301

2.1. Yellow marking mean that the regression (F12) or the independent variables were significant (e.g. 'Seat row in class') within the 95 % Confidence interval, that is <0.05.

2.2. Since the significance of the first regression is already very low, the model appears to be already good. However, there are still many independent variables included with large P-values meaning that they do not predict the 'Age' variable well and introduce noise.

2.3. Therefore, the regression was repeated with only the significant variables.

3. The results of the second regression looked like this

	A	B	C	D	E	F	G	H	I	J	K
10	ANOVA										
11		*df*	*SS*	*MS*	*F*	*Significance F*					
12	Regression	3	781.1772741	260.392425	12.5684344	3.76757E-07					
13	Residual	113	2341.130418	20.7179683							
14	Total	116	3122.307692								
15											
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*	*Age of 6290*	*Rounded Age*
17	Intercept	16.65067896	1.115150029	14.9313353	1.6841E-28	14.44136562	18.85999229	14.4413656	18.8599923		
18	Seat row in class	0.354572821	0.156766179	2.26179411	0.02562292	0.043990745	0.665154896	0.04399075	0.6651549	1	
19	Number of older siblings	1.648337259	0.392745186	4.19696363	5.4119E-05	0.870238211	2.426436307	0.87023821	2.42643631	2	
20	Number of younger siblings	1.682955622	0.457475693	3.67878698	0.0003605	0.776613764	2.58929748	0.77661376	2.58929748	0	
21										20.3019263	20
22											
23										=B17+B18*J18+B19*J19+B20*J20	

3.1. The significance F of the model is even lower; the predictive power is even greater. Therefore, this model will be used instead of the first.

3.2. All P-values for the independent variables are below 0.05 as well.

3.3. To calculate the age of ID 6290, the values of the independent variables are listed in column 'J'. Based on this regression, the age of ID 6290 will be estimated using the values of the independent variables and their respective coefficient. All those relevant values are marked in green (Not entirely visible due to the same highlighted values for the calculation). These values of ID 6290 will be multiplied by the respective coefficient of the independent variables (Column B) and summed with the intercept coefficient (B17). The complete formula is visible in J23. The format of the result was then rounded to whole numbers like the other age values in the demographics sheet. **The age of ID 6290 is thus estimated at roughly 20** (marked in blue).

3.4. The estimated age of ID 6290 will then be transferred back to the demographics sheet. Consequently, the values of ID 6290 will be used for the imputation of other missing values

4. For the other missing values, the process was similar. Therefore they will not be outlined to the same level of detail. The color highlights are the same. Differences and complications in each of the remaining four imputations will be highlighted.

Imputation of the CAO Points of ID 5262

1. First Regression Input
   1.1. Input Y-Range: CAO-Points
   1.2. Input X-Range: All numeric columns expect ID & CAO Points

2. First Regression Results:

	A	B	C	D	E	F	G	H	I
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	10	89854.44662	8985.444662	1.634962326	0.10635798			
13	Residual	107	588051.8246	5495.811445					
14	Total	117	677906.2712						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	418.207914	171.5495942	2.437825142	0.016423332	78.1308512	758.2849768	78.1308512	758.2849768
18	Age (in years)	-2.775911943	1.566477572	-1.772072574	0.079228886	-5.881271031	0.329447146	-5.881271031	0.329447146
19	Daily travel to DCU (in km	0.358010712	0.470858641	0.760335865	0.448726018	-0.575411627	1.291433051	-0.575411627	1.291433051
20	Average year 1 exam resu	2.065199261	0.820924078	2.515700678	0.013365123	0.43781302	3.692585502	0.43781302	3.692585502
21	Seat row in class	5.171116253	2.707084379	1.910216132	0.058782507	-0.195362802	10.53759531	-0.195362802	10.53759531
22	Number of older siblings	-3.611388825	7.008890421	-0.51525828	0.607435437	-17.50569664	10.28291899	-17.50569664	10.28291899
23	Number of younger siblin	-3.785259865	8.030715753	-0.471347758	0.638352056	-19.70521722	12.13469749	-19.70521722	12.13469749
24	Old Dublin postcode (0 if	0.36131797	1.168412245	0.309238431	0.757741523	-1.954923033	2.677558973	-1.954923033	2.677558973
25	Height (in cm)	0.251949989	1.201725136	0.209656918	0.834334122	-2.130329937	2.634229915	-2.130329937	2.634229915
26	Weight (in kg)	-1.134031628	0.885330094	-1.280913904	0.20299328	-2.889095283	0.621032028	-2.889095283	0.621032028
27	Shoe size	0.222868796	4.373086364	0.050963731	0.95944944	-8.446264888	8.892002481	-8.446264888	8.892002481

   2.1. Significance F of Regression still above 0.05 (F12), therefore, the regression should be repeated but with fewer insignificant independent variables.
   2.2. From the independent variables, only ‘Average year 1 results’ is significant. However, one variable might not explain enough variance. Therefore two regressions will follow. One regression only including ‘Average year 1 results’ and another regression including the near-significant variables of ‘Age’ and ‘Seat row’ as well.

3. Result of regression with ‘Average year 1 results’

	A	B	C	D	E	F	G	H	I
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	1	23548.02	23548.02	4.174425914	0.043304318			
13	Residual	116	654358.3	5641.019					
14	Total	117	677906.3						
15									
16		*Coefficients*	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
17	Intercept	375.7239	55.97861	6.711919	7.43455E-10	264.8512193	486.5966	264.8512	486.5966
18	Average year 1 exam result (as %)	1.654528	0.809796	2.043141	0.043304318	0.050624569	3.258432	0.050625	3.258432

   3.1. Both the regression and the independent variables are significant.

4. Result of regression with ‘Average year 1 results’, ‘Age’ and ‘Seat row’

	A	B	C	D	E	F	G	H	I	J	K
9											
10	ANOVA										
11		*df*	*SS*	*MS*	*F*	*Significance F*					
12	Regression	3	70144.17	23381.39	4.385726	0.00583737					
13	Residual	114	607762.1	5331.247							
14	Total	117	677906.3								
15											
16		*Coefficients*	*andard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*	*CAO of 5262*	*Rounded CAO*
17	Intercept	396.8672294	61.75309	6.426678	3.14E-09	274.53484	519.1996	274.5348	519.1996		
18	Age (in years)	-3.449818794	1.359604	-2.53737	0.01252	-6.14318298	-0.75645	-6.14318	-0.75645	21	
19	Average year 1 exam result (as %)	2.010126462	0.797442	2.520718	0.013094	0.43040018	3.589853	0.4304	3.589853	66	
20	Seat row in class	5.499286708	2.559113	2.148903	0.033757	0.42970349	10.56887	0.429703	10.56887	4	
21										479.086528	**479**
22											
23										=B17+B18*J18+B19*J19+B20*J20	

4.1. The Significance F of this regression is even lower and all independent variables are significant. Therefore, **this regression is chosen for the imputation.**

4.2. **The estimated roughly CAO points of ID 5262 is thus 479**. Consequently, the values of ID 5262 will be used for the imputation of other missing values.

Imputation of the Average Year 1 Results Points of ID 7677

1. First Regression Input
   1.1. Input Y-Range: Average Year 1 Results
   1.2. Input X-Range: All numeric columns expect ID & Average Year 1 Results

2. First Regression Results:

	A	B	C	D	E	F	G	H	I
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*gnificance F*			
12	Regression	10	887.5742	88.75742	1.241483818	0.273252			
13	Residual	108	7721.245	71.49301					
14	Total	118	8608.819						
15									
16		*Coefficients*	*andard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
17	Intercept	33.7142	19.76605	1.705662	0.090945593	-5.46555	72.89394	-5.46555	72.89394
18	CAO Points (100 to 600)	0.026938	0.010712	2.514669	0.013387991	0.005704	0.048171	0.005704	0.048171
19	Age (in years)	0.219934	0.179946	1.222221	0.224284974	-0.13675	0.576619	-0.13675	0.576619
20	Seat row in class	-0.5917	0.307973	-1.92126	0.057334769	-1.20215	0.018759	-1.20215	0.018759
21	Daily travel to DCU (in km, 0 if on-campus)	0.037227	0.053719	0.692988	0.48980476	-0.06925	0.143707	-0.06925	0.143707
22	Number of older siblings	-0.14541	0.798493	-0.1821	0.85584544	-1.72816	1.437347	-1.72816	1.437347
23	Number of younger siblings	-0.01883	0.91379	-0.02061	0.983595876	-1.83012	1.792458	-1.83012	1.792458
24	Old Dublin postcode (0 if outside Dublin)	-0.04157	0.13266	-0.31338	0.754595726	-0.30453	0.221382	-0.30453	0.221382
25	Height (in cm)	0.107956	0.135944	0.794121	0.428866628	-0.16151	0.377421	-0.16151	0.377421
26	Weight (in kg)	0.058807	0.099339	0.591981	0.555100517	-0.1381	0.255714	-0.1381	0.255714
27	Shoe size	-0.37426	0.497171	-0.75278	0.453217722	-1.35974	0.611217	-1.35974	0.611217

2.1. Significance F of Regression still above 0.05 (F12), therefore, the regression should be repeated but with fewer insignificant independent variables.

2.2. From the independent variables, only ‘CAO Points’ is significant. However, one variable might not enough variance. Therefore two regressions will follow. One regression only including ‘CAO Points’ and another regression including the near-significant variable of ‘Seat row’.

3. Result of regression with ‘CAO Points’

	A	B	C	D	E	F	G	H	I
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	1	299.8675	299.8675	4.222493179	0.042118212			
13	Residual	117	8308.952	71.01668					
14	Total	118	8608.819						
15									
16		*Coefficients*	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
17	Intercept	58.28896	5.065251	11.50761	5.87625E-21	48.25749318	68.32042	48.25749	68.32042
18	CAO Points (100 to 600)	0.02103	0.010234	2.054871	0.042118212	0.000761674	0.041299	0.000762	0.041299

3.1. Both the regression and the independent variables are significant.

4. Result of regression with ‘CAO Points’ and ‘Seat row’

	A	B	C	D	E	F	G	H	I	J	K
10	ANOVA										
11		*df*	*SS*	*MS*	*F*	*gnificance F*					
12	Regressior	2	473.7649	236.8824	3.37777232	0.037513					
13	Residual	116	8135.054	70.12978							
14	Total	118	8608.819								
15											
16		*Coefficients*	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*	*Grade of 7677*	*ounded Grade*
17	Intercept	59.92061	5.139066	11.65982	2.89625E-21	49.74204	70.09918	49.74204	70.09918		
18	CAO Point	0.022823	0.010234	2.230183	0.027659479	0.002554	0.043093	0.002554	0.043093	600	
19	Seat row in	-0.44358	0.281692	-1.57469	0.118051303	-1.00151	0.114349	-1.00151	0.114349	6	
20										70.95311832	**71**
21											
22										=B17+B18*J18+B19*J19	

4.1. The Significance F of this regression is even lower and all independent variables are significant except for ‘Seat row’. However, since the overall predictive power of the model is slightly better, **this regression is chosen for the imputation.**

4.2. The estimated rough Average **First Year Grade of ID 7677 is 71.** Consequently, the values of ID 7677 will be used for the imputation of other missing values

Imputation of the Daily travel to DCU of ID 7181

1. First Regression Input
   1.1. Input Y-Range: Daily Travel to DCU
   1.2. Input X-Range: All numeric columns expect ID & Daily Travel to DCU

2. First Regression Results:

	A	B	C	D	E	F	G	H	I
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	10	3298.139	329.8139	1.448372917	0.168995402			
13	Residual	109	24820.76	227.7134					
14	Total	119	28118.9						
15									
16		*Coefficients*	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
17	Intercept	-70.261	34.28151	-2.04953	0.042810771	-138.2058807	-2.31621	-138.206	-2.31621
18	Average year 1 exam result (as	0.117469	0.171342	0.685581	0.494432918	-0.2221258	0.457064	-0.22213	0.457064
19	CAO Points (100 to 600)	0.016745	0.019519	0.857867	0.392848144	-0.021941473	0.055431	-0.02194	0.055431
20	Seat row in class	0.660111	0.555279	1.18879	0.237105707	-0.440435076	1.760657	-0.44044	1.760657
21	Age (in years)	0.213074	0.322363	0.660974	0.510023364	-0.425839256	0.851987	-0.42584	0.851987
22	Number of older siblings	2.346644	1.407392	1.66737	0.098311534	-0.442761802	5.13605	-0.44276	5.13605
23	Number of younger siblings	2.667785	1.593121	1.674565	0.096886539	-0.489729112	5.8253	-0.48973	5.8253
24	Old Dublin postcode (0 if outsi	-0.05054	0.236167	-0.21399	0.830955931	-0.518612916	0.417539	-0.51861	0.417539
25	Height (in cm)	0.339041	0.233465	1.452215	0.149314301	-0.123678441	0.80176	-0.12368	0.80176
26	Weight (in kg)	-0.10374	0.174507	-0.59449	0.553414678	-0.449611484	0.242125	-0.44961	0.242125
27	Shoe size	-0.20676	0.877386	-0.23565	0.814144623	-1.945709714	1.532193	-1.94571	1.532193

   2.1. Significance F of Regression is still above 0.05 (F12), therefore, the regression should be repeated but without the very insignificant independent variables.
   2.2. From the independent variables none are significant. Therefore two regressions will follow. One regression will include 'Number of older siblings' and 'Number of younger siblings' as those were nearest to 0.05. Another regression will include all the near-significant variables of 'Number of older siblings', 'Number of younger siblings', 'Seat row' and 'Height'.

3. Result of regression with 'Number of older siblings', 'Number of younger siblings', 'Seat row' and 'Height'.

	A	B	C	D	E	F	G	H	I
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*gnificance F*			
12	Regression	4	2685.866	671.4665	3.036156	0.020202			
13	Residual	115	25433.03	221.1568					
14	Total	119	28118.9						
15									
16		*Coefficients*	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
17	Intercept	-38.2031	24.50126	-1.55923	0.12169	-86.7354	10.32918	-86.7354	10.32918
18	Seat row in class	0.760712	0.509178	1.493999	0.137915	-0.24787	1.769296	-0.24787	1.769296
19	Number of older siblings	2.561834	1.274715	2.009731	0.046801	0.036869	5.0868	0.036869	5.0868
20	Number of younger siblings	2.783837	1.475419	1.886811	0.061707	-0.13868	5.706359	-0.13868	5.706359
21	Height (in cm)	0.217156	0.135152	1.606754	0.110851	-0.05055	0.484865	-0.05055	0.484865

   3.1. Regression significance is within the confidence interval.
   3.2. However, only 'Number of siblings' has a lower P-value than 0.05.

4. Result of regression with Number of older siblings' and 'Number of younger siblings'

	A	B	C	D	E	F	G	H	I	J	K
10	ANOVA										
11		*df*	*SS*	*MS*	*F*	*gnificance F*					
12	Regression	2	1745.209	872.6047	3.871084	0.023556					
13	Residual	117	26373.69	225.4161							
14	Total	119	28118.9								
15											
16		*Coefficients*	*andard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*	*Travel of 7677*	*Rounded Travel*
17	Intercept	3.964749	2.60526	1.521825	0.130751	-1.19483	9.124331	-1.19483	9.124331		
18	Number of older siblings	2.822662	1.269081	2.224178	0.028056	0.309313	5.336011	0.309313	5.336011	0	
19	Number of younger siblings	2.902576	1.485585	1.953826	0.053108	-0.03955	5.844699	-0.03955	5.844699	1	
20										6.8673251	7
21											
22										=B17+B18*J18+B19*J19	

4.1. The Significance F of this regression is slightly higher all independent variables are not significant except for 'Number of older siblings'.

4.2. This regression, compared to the former, is simpler with fewer variables, and both independent variables are significant or near to it. The former regression includes two additional predictors, but neither adds statistical significance to the model. Adding complexity without a clear increase in explanatory power can be unnecessary. Furthermore, simplicity is preferred in model selection because it reduces the risk of overfitting and makes the model easier to interpret **this model was chosen.**

4.3. **The estimated Travel to DCU in km of ID 7677 is 7.** Consequently, the values of ID 7677 will be used for the imputation of other missing values.

Imputation of the Seat Row in Class of ID 0838

1. First Regression Input
   1.1. Input Y-Range: Seat Row
   1.2. Input X-Range: All numeric columns expect ID & Seat Row

2. First Regression Results:

	A	B	C	D	E	F	G	H	I
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	10	167.5355	16.75355	2.527134666	0.008876539			
13	Residual	110	729.2413	6.629466					
14	Total	120	896.7769						
15									
16		*Coefficients*	*andard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
17	Intercept	6.20433	5.887926	1.053738	0.294312102	-5.464157736	17.87282	-5.46416	17.87282
18	Daily travel to DCU (in km, 0 if on-campus)	0.01944	0.016235	1.197378	0.233733285	-0.012734594	0.051614	-0.01273	0.051614
19	Average year 1 exam result (as %)	-0.05574	0.028809	-1.93463	0.05560377	-0.11282919	0.001358	-0.11283	0.001358
20	CAO Points (100 to 600)	0.006279	0.003288	1.909879	0.058753153	-0.000236327	0.012794	-0.00024	0.012794
21	Age (in years)	0.141076	0.053407	2.641531	0.009455047	0.03523605	0.246916	0.035236	0.246916
22	Number of older siblings	0.083966	0.241304	0.347968	0.728529373	-0.394241054	0.562173	-0.39424	0.562173
23	Number of younger siblings	-0.07343	0.275114	-0.26692	0.790033093	-0.618644009	0.471779	-0.61864	0.471779
24	Height (in cm)	-0.01927	0.039961	-0.4822	0.63062088	-0.098462834	0.059924	-0.09846	0.059924
25	Old Dublin postcode (0 if outside Dublin)	0.000373	0.040286	0.00927	0.992620201	-0.079463035	0.08021	-0.07946	0.08021
26	Weight (in kg)	0.036202	0.029609	1.22268	0.224063784	-0.022475428	0.094879	-0.02248	0.094879
27	Shoe size	-0.28847	0.147046	-1.96175	0.052318666	-0.579875896	0.002944	-0.57988	0.002944

2.1. Significance F of Regression is already slightly below 0.05 (F12). However, the regression should be repeated but with fewer very insignificant independent variables.

2.2. From the independent variables, only 'Age' is significant. However, one variable might not explain enough variance. Therefore two regressions will follow. One

regression only including ‘Age’ and almos significant ‘Shoe Size’ and another regression also including ‘Average year 1 reults’ and ‘CAO Points’.

3. Result of regression with ‘Age’, ‘Shoe Size’, ‘Average year 1 reults’ and ‘CAO Points’

	A	B	C	D	E	F	G	H	I
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*'gnificance F*			
12	Regression	4	145.8577	36.46441	5.632926	0.000352			
13	Residual	116	750.9192	6.473441					
14	Total	120	896.7769						
15									
16		*Coefficients*	*andard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
17	Intercept	4.440898	2.631885	1.687345	0.094224	-0.77188	9.653679	-0.77188	9.653679
18	Average year 1 exam result (as %)	-0.05185	0.028156	-1.84159	0.06809	-0.10762	0.003915	-0.10762	0.003915
19	CAO Points (100 to 600)	0.006144	0.003202	1.918833	0.057463	-0.0002	0.012486	-0.0002	0.012486
20	Age (in years)	0.164376	0.046367	3.545107	0.000567	0.072541	0.256212	0.072541	0.256212
21	Shoe size	-0.22756	0.097425	-2.33579	0.021219	-0.42053	-0.0346	-0.42053	-0.0346

3.1. Both the regression and the independent variables of ‘Age’ and ‘Shoe size’ are significant; thus, underlining the idea to just use those two variables for a regression.

4. Result of regression with ‘Age’ and ‘Shoe Size’

	A	B	C	D	E	F	G	H	I	J	K
9											
10	ANOVA										
11		*df*	*SS*	*MS*	*F*	*'gnificance F*					
12	Regression	2	108.058	54.02902	8.083266	0.000513					
13	Residual	118	788.7188	6.684058							
14	Total	120	896.7769								
15											
16		*Coefficients*	*andard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Upper 95.0%*	*Row of 0838*	*Rounded Row*
17	Intercept	4.628397	1.304146	3.548986	0.000556	2.045832	7.210962	2.045832	7.21096157		
18	Age (in years)	0.143686	0.046214	3.109166	0.002352	0.05217	0.235201	0.05217	0.235201249	19	
19	Shoe size	-0.26064	0.097953	-2.66089	0.008879	-0.45461	-0.06667	-0.45461	-0.066667872	9	
20										5.01265794	5
21											
22										=B17+B18*J18+B19*J19	

4.1. The Significance F of this regression is slightly higher; however, all independent variables are significant except. **Therefore, this regression is chosen for the imputation**

4.2. The estimated Seat Row of ID 0838 is 5.

Personality Table

1. Duplicates
    a. Duplicate ID 1699:
    Entry ID 1699 has two entries in the personalities sheet. Since those entries are identical, one is deleted.
    b. Duplicate ID 4397:
    The duplicates of ID 4397 were both removed. Although it is highly unfortunate to lose entries for an already small dataset, it is impossible to ascertain which, if any, of the conflicting records represent the true values, retaining either could introduce bias or error into the dataset. This is particularly critical for the ensuing data analysis process where the cost of incorrect information could be significant. Especially with such a small dataset, removing both duplicates ensures that any insights or models derived from the data are not skewed by these uncertainties. Thus entries of ID 4397 were removed in the Personality sheet.
    c. Duplicate ID 1462:
    Same argumentation as with ID 1462. Both entries were removed from the personality sheet.
2. Low Personality scores ID 2785:
    This entry is indeed interesting. One could argue that this is a valid entry as such value is possible in the personality test and therefore should be kept. However, several arguments speak for removing the entry. First, if one of those personality dimensions is very low or very high, this would appear more realistic. However, a person scoring extremely low in all five traits in such a small dataset seems highly unlikely. It cannot be rooted out that this may be a valid entry as those values are possible. However, it appears even more likely that somebody made a mistake when entering the results or wanted to "troll" the rest of the class. Therefore, also this ID is removed from the personality tab even if this may introduce potential bias. This leaves 97 unique entries in the personality tab.
3. Merging:
    The demographics data was merged with the personality data in Excel. The last four digits of the phone number were utilized as the key to align the rows between the two datasets. The VLOOKUP function was employed to search for and retrieve the corresponding personality information for each entry in the demographics dataset. Rows in the demographics data without a corresponding entry in the personality dataset were left with blank spaces, as the personality data had fewer rows. Ultimately, a unified dataset was created, displaying combined demographics and personality data wherever a match was found. Some entries from the demographics data were left without personality information due to the absence of matching phone number digits.
    Since I did not work with the actual "Merged" Sheet, I have inserted a screenshot on the next page to underline that I still merged the tables using this Vlook-up function.

Now that the sample has been cleaned and missing values are imputed, the next section will focus on the Data Analytics report itself.

Q2

=VLOOKUP($A2; Table2[#All];2;FALSE)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Last 4 digits of your mobile (0000 to 9999)	Seat row in class	Daily travel to DCU (in km, 0 if on-campus)	CAO Points (100 to 600)	Average year 1 exam result (as %)	Age (in years)	Shoe size	Number of older siblings	Number of younger siblings	Height (in cm)	Old Dublin postcode (0 if outside Dublin)	Weight (in kg)	Gender	Eye colour	Hair colour	Star sign	PT_EXTRAVERSION	PT_INTUITION	PT_THINKING	PT_JUDGING	PT_ASSERTIVE
2	0059	10	1	510	70	20	8	0	0	160	0	76	Male	Brown	Black	Gemini	#N/A	#N/A	#N/A	#N/A	#N/A
3	0087	10	2	566	70	19	5	2	0	167	0	60	Female	Green	Brown	Pices	#N/A	#N/A	#N/A	#N/A	#N/A
4	0111	5	4.5	500	63	22	10	1	0	167	7	51	Female	Green	Blonde	Aquarius	48	49	39	67	26
5	0130	7	9	513	61	19	10	1	0	170	3	79	Male	Blue	Black	Gemini	#N/A	#N/A	#N/A	#N/A	#N/A
6	0460	3	31	555	80	20	10	1	0	181	0	80	Male	Brown	Brown	Taurus	60	49	40	65	69
7	0602	8	5.5	525	70	22	5	1	0	165	0	56	Female	Blue	Brown	Gemini	#N/A	#N/A	#N/A	#N/A	#N/A
8	0608	3	10	509	64	19	10	0	0	187	3	68	Male	Brown	Brown	Aquarius	#N/A	#N/A	#N/A	#N/A	#N/A
9	0730	5	7	450	65	20	8	1	1	174	7	62	Female	Green	Black	Aquarius	81	61	72	39	33
10	0817	3	0.05	435	66	22	10.5	0	2	158	0	69	Male	Blue	Blonde	Gemini	#N/A	#N/A	#N/A	#N/A	#N/A
11	0838	5	10	543	71	19	9	0	1	187	13	76	Male	Blue	Red	Taurus	48	41	57	49	36
12	0844	3	2	500	66	21	6	0	1	174	0	60	Female	Green	Blonde	Leo	#N/A	#N/A	#N/A	#N/A	#N/A
13	0904	5	3	357	42	23	11	2	0	187	0	75	Male	Blue	Brown	Aquarius	56	73	53	47	37
14	0937	3	3	588	70	22	9	1	0	190	11	82	Male	Blue	Blonde	Taurus	#N/A	#N/A	#N/A	#N/A	#N/A
15	0949	3	1	479	86	19	11	2	0	187	9	87	Male	Blue	Brown	Virgo	63	42	43	42	49
16	0971	8	5	567	68	20	6	4	0	169	9	63	Female	Blue	Black	Sagittarius	46	48	34	64	42
17	0987	4	10	552	80	20	8	3	0	170	3	58	Female	Green	Red	Gemini	66	65	40	53	29
18	0999	2	4.7	550	73	21	13	1	0	193	1	88	Male	Brown	Brown	Gemini	54	49	58	40	58
19	1024	6	7	450	66	21	8	1	2	176	7	80	Male	Brown	Brown	Leo	#N/A	#N/A	#N/A	#N/A	#N/A
20	1049	7	5	450	57	18	5	0	1	163	9	56	Female	Brown	Brown	Libra	64	59	49	63	26
21	1128	1	0	500	90	19	7	0	0	168	3	66	Female	Blue	Red	Scorpio	56	56	71	63	64
22	1145	6	5	552	68	20	10	1	1	177	11	70	Male	Green	Brown	Aquarius	#N/A	#N/A	#N/A	#N/A	#N/A
23	1238	6	12	400	55	22	8	1	0	173	0	58	Male	Blue	Brown	Libra	#N/A	#N/A	#N/A	#N/A	#N/A
24	1278	2	1	504	76	21	9	1	1	175	0	82	Male	Blue	Black	Leo	73	66	33	33	22
25	1291	6	0	507	70	20	5.5	0	0	163	0	57	Female	Green	Brown	Pices	#N/A	#N/A	#N/A	#N/A	#N/A
26	1352	7	20	355	66	22	13	0	2	193	0	97	Male	Blue	Brown	Libra	#N/A	#N/A	#N/A	#N/A	#N/A
27	1446	11	4.8	510	63	19	6	0	0	162	7	62	Female	Brown	Brown	Pices	39	61	33	21	11
28	1462	6	5	500	80	21	8	1	2	169	12	57	Female	Green	Black	Leo	#N/A	#N/A	#N/A	#N/A	#N/A
29	1546	3	7	500	75	24	6.5	0	3	174	1	70	Female	Green	Blonde	Aquarius	#N/A	#N/A	#N/A	#N/A	#N/A
30	1699	3	7	565	71	21	6	2	1	170	14	59	Female	Blue	Black	Libra	64	69	63	58	60
31	1816	6	0	496	70	20	9	0	1	182	0	80	Male	Blue	Brown	Aquarius	46	62	72	21	60
32	2228	12	2	550	75	23	7	1	0	175	1	75	Male	Brown	Brown	Capricorn	74	51	52	68	83
33	2231	5	46	400	63	22	10	2	1	182	12	77	Male	Blue	Blonde	Sagittarius	#N/A	#N/A	#N/A	#N/A	#N/A
34	2288	10	4	498	72	19	6	1	1	165	11	60	Female	Brown	Black	Virgo	42	72	62	61	31
35	2356	8	6.7	550	99	33	10	2	1	180	12	80	Male	Green	Black	Scorpio	#N/A	#N/A	#N/A	#N/A	#N/A
36	2408	6	20	524	70	21	10	1	1	170	7	68	Female	Green	Brown	Sagittarius	#N/A	#N/A	#N/A	#N/A	#N/A
37	2424	3	3	490	73	26	8	0	2	169	0	60	Female	Brown	Brown	Gemini	46	43	66	47	49
38	2523	9	1.7	521	70	20	10	0	1	189	11	68	Male	Blue	Brown	Libra	#N/A	#N/A	#N/A	#N/A	#N/A
39	2561	5	17	554	76	20	8	1	2	183	15	80	Male	Green	Brown	Sagittarius	#N/A	#N/A	#N/A	#N/A	#N/A
40	2576	3	3	468	60	20	12	1	1	198	11	90	Male	Green	Brown	Capricorn	#N/A	#N/A	#N/A	#N/A	#N/A
41	2666	4	78	578	72	20	8.5	3	0	180	0	63	Male	Green	Brown	Capricorn	52	69	49	67	42
42	2785	4	5	521	68	20	9	1	0	178	9	77	Male	Blue	Red	Libra	#N/A	#N/A	#N/A	#N/A	#N/A
43	2803	5	3.5	400	48	21	10	1	0	180	0	75	Male	Brown	Brown	Aries	74	42	36	68	18
44	2955	3	15	460	70	21	6	1	1	167	15	79	Female	Green	Brown	Leo	#N/A	#N/A	#N/A	#N/A	#N/A
45	2986	5	9	498	65	19	5.5	1	2	156	13	55	Female	Blue	Brown	Cancer	64	49	36	39	51
46	3173	5	5	425	65	21	6.5	1	1	175	0	63	Female	Blue	Blonde	Scorpio	24	42	29	64	30
47	3220	5	1	555	68	21	11	2	1	188	9	84	Male	Brown	Black	Sagittarius	#N/A	#N/A	#N/A	#N/A	#N/A

Blueprints of the Mind: The CA259 Class Personality Survey

To ensure a concise and focused data analysis, it is crucial to select a dataset that promises valuable insights. Among the three available sheets, namely 'Personality', 'Demographics', and 'Merged', a focused approach is advisable, particularly for a brief report. Although the 'Demographics' sheet is rich in data, its potential patterns have been partially explored during the process of imputing missing values, reducing the novelty of any findings in this report. However, the 'Merged' sheet has a smaller sample size of 62 entries, limiting the ability to explore correlations between demographic characteristics and personality traits. As a result, no robust conclusions can be drawn from its analysis. Therefore, the 'Personality' sheet is the most promising candidate for this exercise. This report aims to analyze the personality profiles within the CA259 class and compare them to broader Irish and European populations. The focus is on illuminating the distinct personalities (McDermott & Spann, 2024) present in the CA259 class and providing meaningful context to their distribution and characteristics.

The standout feature in the CA259 class personality survey is the high average Judging score of 57.45%, suggesting a class preference for structure and organization. This aligns well with the demands of the CA259 Data Analytics for Marketing Applications course, which likely requires students to be decisive and systematic, qualities that may resonate with both the second-year MINT and final-year Global Business students involved. However, the Assertive trait shows a different pattern, with the largest standard deviation of 19.08% and a mean score near the midpoint of the scale at 49.37%. This implies a wide variation in the confidence levels with which students approach their lives and work. In the 16 Personalities framework, Assertiveness relates to one's confidence in their abilities and decisions, often impacting their stress levels and the way they respond to challenges. The substantial spread in assertiveness might reflect the varying experiences and backgrounds among the students. For example, the broad range in Assertiveness is particularly intriguing when considering the mix of MINT and Global Business students. It may have been expected that these students would demonstrate higher levels of assertiveness due to the demanding nature of their fields, which often require leadership and self-assurance. The other personality traits (Extraversion, Intuition, and Thinking) show less variation in their means and standard deviations, suggesting a more moderate distribution of these characteristics within the class. The mean scores for these traits are also closer to the midpoint of the possible range, indicating no strong lean towards either pole of these personality dimensions.

	N	Min.	Max.	Mean	Median	SD	Range
Extraversion	97	17	93	55.32	59	18.25	76
Intuition	97	23	93	53.65	52	13.07	70
Thinking	97	19	82	48.94	49	13.84	63
Judging	97	21	94	57.45	58	17.14	73
Assertive	97	8	93	49.37	49	19.08	85

Figure 1 Descriptive Statistics of 'Personality' Sheet

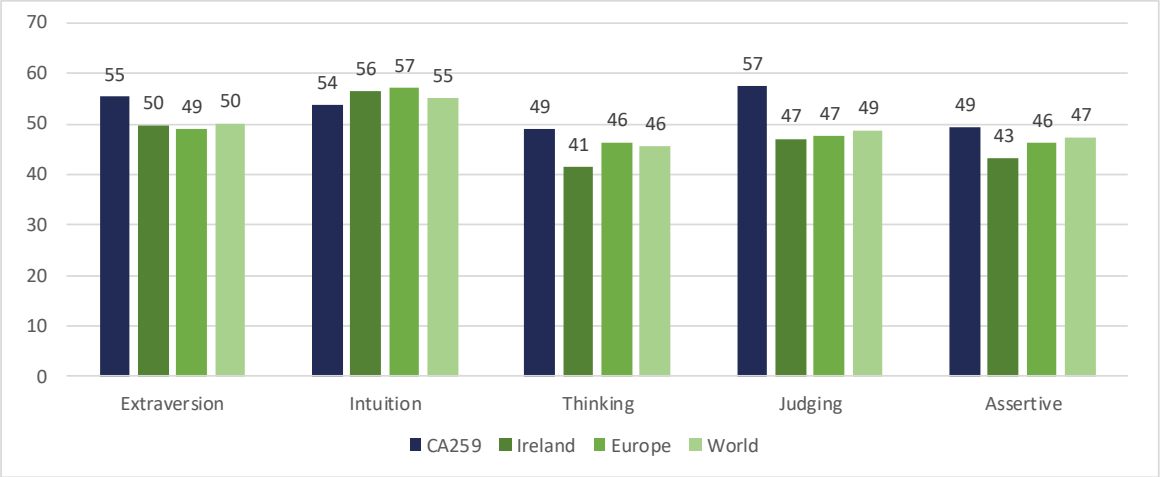


Figure 2 - Personality Dimensions of CA259, Ireland, Europe and World (Ireland Personality Profile, n.d.)

Comparing the CA259 to larger populations, the class shows a higher degree of Extraversion (55.32%) compared to Ireland, Europe, and the global average, indicating a more outgoing and energetic temperament among these students. Intuition is also marginally higher in CA259 (53.65%) than in Ireland and the World, but just below Europe's average. This slight difference suggests that the class is inclined towards abstract thinking and future possibilities, which is consistent with the European average. Thinking, representing logical decision-making, is higher in CA259 (48.94%) than in Ireland and slightly above the World average, but still below the European figure. This could reflect the influence of the European students in the class who might bring a more analytical approach to the group. Judging in CA259 (57.45%) is notably higher than Ireland, Europe, and the World. This trait, indicative of a preference for structure and decisiveness, is strongest within the class, possibly due to the educational environment which often requires planning and organization. Assertiveness in CA259 (49.37%) is just above the Irish and European averages but below the World average, suggesting a moderate level of confidence and even-temperedness among the students. Thus, in comparison to larger populations, CA259 stands out for its high Extraversion, strong Judging preference, and moderate Assertiveness, reflecting its diverse and structured educational context.

Regarding personality types, the MBTI scores for the class of CA259 were categorised using a threshold method. Scores under 50 were allocated to one side of the MBTI scale (**E**xtraversion, i**N**tuition, **T**hinking, **J**udgement, **A**ssertiveness), while scores of 50 or more were assigned to the other (**I**ntroversion, **S**ensing, **F**eeling, **P**erception, **T**urbulence). As a result, these classifications were combined to create a four-letter MBTI personality code for each individual. The Personality codes were then grouped into the Personality types of “Analyst”, “Diplomat”, “Explorer” and “Sentinel” (Our Framework, n.d.) to find out whether there are many Analysts in this Data Analytics course.

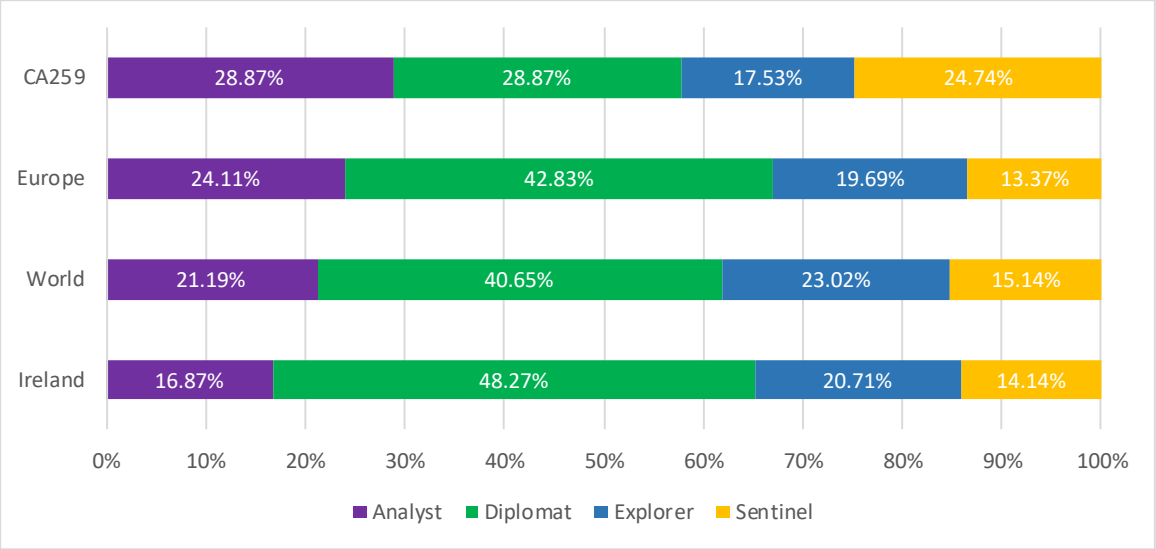


Figure 3 - Personalities of CA259, Ireland, Europe and World (Ireland Personality Profile, n.d.)

The personality distribution in the CA259 class reveals some interesting patterns when compared with larger populations. At 28.87%, Analysts are substantially overrepresented in CA259 compared to Ireland, Europe, and the world. This could align with the analytical nature of the Data Analytics for Marketing Applications course, which may naturally attract individuals with a preference for logical and strategic thinking, typical of the Analyst type. Diplomats, on the other hand, are significantly underrepresented in CA259 (28.87%) relative to Ireland's, Europe's, and the world's scores. This might suggest that the interpersonal and empathetic skills associated with Diplomats are less drawn to this course's technical focus. Explorers are slightly less common in CA259 (17.53%) than in the global and Irish populations, but more in line with Europe. Explorers are known for their spontaneity and flexibility, traits that may be less central to a curriculum that likely values structured data analysis. Sentinels in CA259 (24.74%) are more prevalent than in all three comparison groups: Ireland, Europe, and the world. This suggests that students with a preference for order, security, and stability are more attracted to or prevalent in the CA259 course.

In summary, the CA259 class has a unique personality profile compared to the wider Irish and European populations. They display a strong preference for Judging, indicating a leaning towards structure and organization. This trait is advantageous given the rigorous nature of their studies in Data Analytics for Marketing Applications. A varied range of Assertiveness levels indicates a dynamic classroom where different levels of confidence may complement each other, fostering a supportive yet challenging environment. The overrepresentation of Analysts and underrepresentation of Diplomats highlight the course's appeal to those with analytical mindsets. The prevalence of Sentinels underscores the value placed on order, mirroring the structured approach necessary for data analysis. Overall, the personality survey seems to align with the personality traits expected from MINT and Global Business students.

List of References

Ireland Personality Profile | Country Personality Profiles | 16Personalities. (n.d.). 16 Personalities. Retrieved 17 February 2024, from https://www.16personalities.com/country-profiles/ireland

McDermott, N., & Spann, R. T. (2024, January 5). *Myers-Briggs Type Indicator (MBTI): A Beginner's Guide*. Forbes Health. https://www.forbes.com/health/mind/myers-briggs-personality-test/

Our Framework | 16Personalities. (n.d.). 16 Personalities. Retrieved 18 February 2024, from https://www.16personalities.com/articles/our-theory