Assignment #1 (of 3): Clean and Report on a Dataset

Academic Assignment Number One

in the study programme [redacted]
and the module [redacted]

Submitted by:
Leonie Kim Röskam
[redacted]

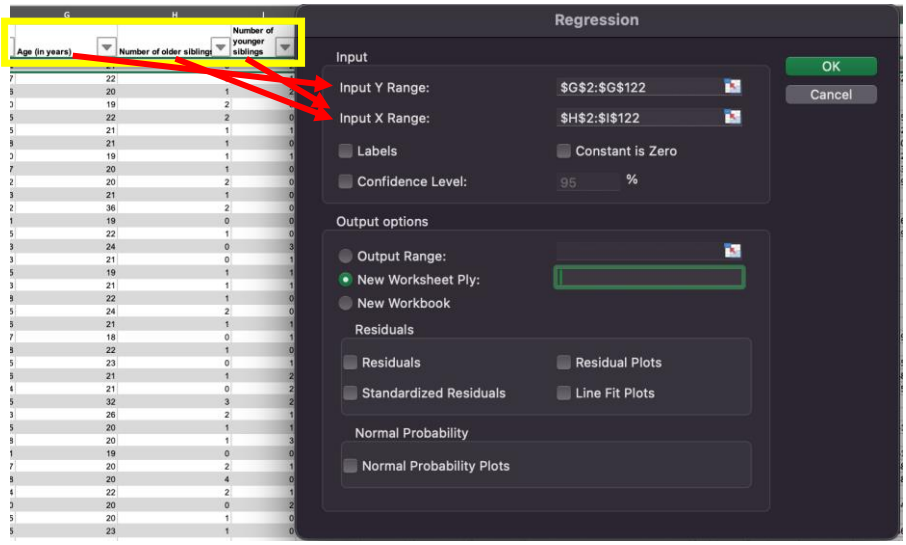
Submitted to:
[redacted]

Dublin, 25.02.2024

Part 1: Normalising the dataset and merging with the Personalities dataset

In the initial stage, we supplemented the incomplete dataset, which was filled in by the results of the survey for an 'imaginary friend', with the five missing values. For this, we used the Regression functionality of Microsoft Excel which was extensively discussed in class. The process on how we calculated the missing values will be described in detail by the example of "Age" and was likewise used to calculate the other missing values for "Seat Row", "Exam Result", "CAO Result" and "Daily Travel to DCU".

1. The first missing value is Age, it is the dependant or intercept variable in this case. We determined older and younger siblings to be the independent variable given that these are the only variables that are to some extent related to age distribution. We started with this value, as there were no missing values for the independent variables in the dataset, thus could use all the rows from the demographics dataset to calculate the missing age (121 rows (Observations in Excel) + 1 row with missing age = 122 rows). We used the Data Analysis add-in tool in Excel. We used the Regression functionality and added the input for the Y range to be the Age (in years) column. For the X range input, we choose the columns Number of older siblings and Number of younger siblings.



The results of the regression were the following:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,463042331
R Square	0,2144082
Adjusted R S	0,201093085
Standard Err	4,571615579
Observations	121

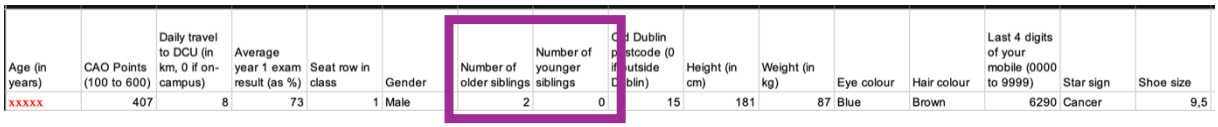
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	673,0787274	336,539364	16,1026169	6,55571E-07
Residual	118	2466,160942	20,899669		
Total	120	3139,239669			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95,0%*	*Upper 95,0%*
Intercept	18,43273903	0,784650008	23,4916	2,5967E-46	16,87891839	19,9865597	16,8789184	19,9865597
X Variable 1	1,803448059	0,382880224	4,710214	6,8129E-06	1,045240977	2,56165514	1,04524098	2,56165514
X Variable 2	1,700254638	0,453863866	3,7461775	0,00027927	0,801480619	2,59902866	0,80148062	2,59902866

In the next step, we analysed the significance F (marked by green border), which was below 0.05, meaning that the regression was significant. We then analysed the p-values of the independent variables (marked in red) which were also both significant as they were below 0.05. Following that, we imputed the missing value with the predictive model we created using the following formula:

Missing value = Intercept coefficient + Coefficient independent variable 1 * Value independent variable 1 + Coefficient independent variable 2 * Value independent variable 2 + Coefficient independent variable n * Value independent variable n



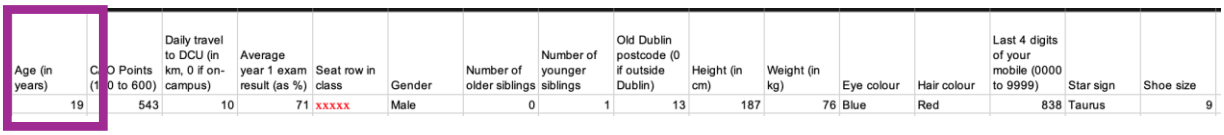
Age (in years)	CAO Points (100 to 600)	Daily travel to DCU (in km, 0 if on-campus)	Average year 1 exam result (as %)	Seat row in class	Gender	Number of older siblings	Number of younger siblings	Old Dublin postcode (0 if outside Dublin)	Height (in cm)	Weight (in kg)	Eye colour	Hair colour	Last 4 digits of your mobile (0000 to 9999)	Star sign	Shoe size
xxxxx	407	8	73	1	Male	2	0	15	181	87	Blue	Brown	6290	Cancer	9,5

The calculation was performed in Excel and looked like this:

Age= 18.43273903 + 1.803448059 * 2 + 1.70025463764142 * 0
Age = 22.0396351

Consequently, the value to be filled in is 22.04, which means the person is predicted to be 22 years old having two older siblings.

2. Secondly, we imputed the missing value for seat row. We tested different combinations of independent variables. Firstly, CAO and exam results which had a significance F of 0.142 therefore being insignificant. Then we analysed distance to DCU and Dublin Postcode which, with 0.233 significance F, was also insignificant, both p-values were also insignificant. Afterwards we tested all four variables from above together which had a significance F of 0.13, so also insignificant. In order to calculate the regression, we moved the corresponding rows with the missing values out of the sheet, adapting it for every regression. Lastly, we tried age as the independent variable, now having the age for all 122 rows, and received a significance F of 0.0053 and continued the imputation with this variable:

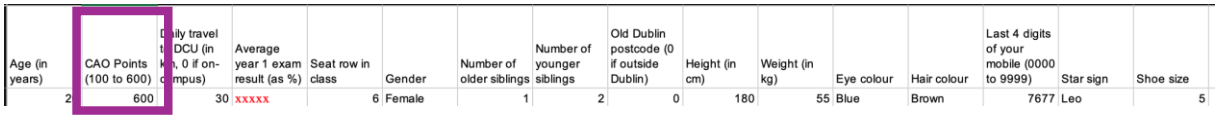


Age (in years)	CAO Points (100 to 600)	Daily travel to DCU (in km, 0 if on-campus)	Average year 1 exam result (as %)	Seat row in class	Gender	Number of older siblings	Number of younger siblings	Old Dublin postcode (0 if outside Dublin)	Height (in cm)	Weight (in kg)	Eye colour	Hair colour	Last 4 digits of your mobile (0000 to 9999)	Star sign	Shoe size
19	543	10	71	xxxxx	Male	0	1	13	187	76	Blue	Red	838	Taurus	9

Seat row = 2.6338454 + 0.13640756 * 19
Seat row = 5.22558897

Consequently, the value to be filled in is 5.22558897 which means the person is predicted to sit in row 5.

3. The third missing value was exam results, which we calculated with CAO results as the independent variable given the connection that both are education related metrics. The regression had a significance f and p-value of 0.0411 and was thus significant. Given that one missing value was also the exam result, we excluded that row from our calculations.

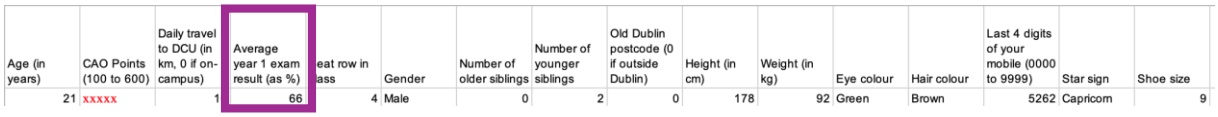


Age (in years)	CAO Points (100 to 600)	Daily travel to DCU (in km, 0 if on-campus)	Average year 1 exam result (as %)	Seat row in class	Gender	Number of older siblings	Number of younger siblings	Old Dublin postcode (0 if outside Dublin)	Height (in cm)	Weight (in kg)	Eye colour	Hair colour	Last 4 digits of your mobile (0000 to 9999)	Star sign	Shoe size
2	600	30	xxxxx	6	Female	1	2	0	180	55	Blue	Brown	7677	Leo	5

Exam result = 58.2882579 + 0.02109449* 600
Exam result = 70.9449548

Consequently, the value to be filled in is 70.94495481 which means the person is predicted to have average exam results of 71.

4. As we now knew all exam results, we then calculated the fourth missing value, CAO result. The first and in our eyes related variable tested was the Year 1 exam results the significance f and the p-value were 0.0378, meaning significant.

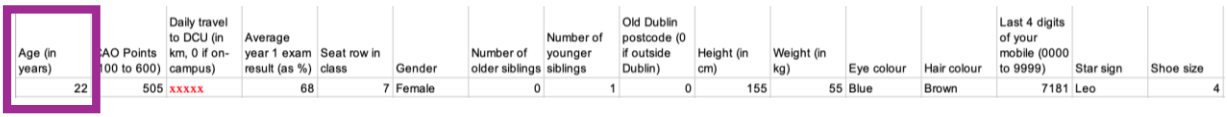


Age (in years)	CAO Points (100 to 600)	Daily travel to DCU (in km, 0 if on-campus)	Average year 1 exam result (as %)	eat row in lass	Gender	Number of older siblings	Number of younger siblings	Old Dublin postcode (0 if outside Dublin)	Height (in cm)	Weight (in kg)	Eye colour	Hair colour	Last 4 digits of your mobile (0000 to 9999)	Star sign	Shoe size
21	xxxxx	1	66	4	Male	0	2	0	178	92	Green	Brown	5262	Capricorn	9

CAO result = 374.2867332 + 1.69634599 * 66
CAO result = 486.245569

Consequently, the value to be filled in is 486.245569 which means the person is predicted to have a CAO result of 486 points.

5. Lastly, we imputed the missing value for Daily Travel. In our eyes there were no independent variables which had a real-life relation to this variable as the only distance related variable would be the Dublin postcode but the numbers within that variable give no information about the actual distance to DCU. This is why we tested different independent variables individually to see if there would be a significant one. The only significant independent variable we found was age with a significance f and p value of 0.0368, so we imputed the final missing value with this:



Age (in years)	AO Points 00 to 600)	Daily travel to DCU (in km, 0 if on-campus)	Average year 1 exam result (as %)	Seat row in class	Gender	Number of older siblings	Number of younger siblings	Old Dublin postcode (0 if outside Dublin)	Height (in cm)	Weight (in kg)	Eye colour	Hair colour	Last 4 digits of your mobile (0000 to 9999)	Star sign	Shoe size
22	505	xxxxx	68	7	Female	0	1	0	155	55	Blue	Brown	7181	Leo	4

Daily Travel = -2.5597364 + 0.5713765 * 22
Daily Travel = 10.0105466

Consequently, the value to be filled in is 10,0105466 which means the person is predicted to have a daily travel distance of 10.01km to get to DCU.

Having calculated the missing values for the demographics dataset, we now moved on to merging the dataset with the Personalities dataset on the 4 digit number. For this, we used Python. The code to merge looked like this:

```python
import pandas as pd

# Load the Excel file
file_path = '/Users/anika/Downloads/CA259 Students.xlsx'
xl = pd.ExcelFile(file_path)

# Load the two sheets into two DataFrames
df1 = xl.parse(0)  # The first sheet is the one to merge into
df2 = xl.parse(1)  # The second sheet contains the data to merge

# Identify the 4-digit number columns in both DataFrames
column_4_digit_1 = df1.columns[13]
column_4_digit_2 = df2.columns[0]

# Merge the DataFrames on the 4-digit number column
merged_df = pd.merge(df1, df2, left_on=column_4_digit_1, right_on=column_4_digit_2, how='left')

merged_df # Display the first few rows of the merged DataFrame

output_file_path = '/Users/anika/Downloads/merged_CA259 Students.xlsx'
merged_df.to_excel(output_file_path, index=False)
```

The VLOOKUP function in Excel could have also been used to complete this task.

The merged dataset contains 65 rows that include complete data. Consequently, it can be assumed that 57 people did either not match their phone number correctly in the surveys or only did one of both surveys. In the next step, we started to explore the merged dataset to analyse the different aspects of the dataset and build a story around the dataset of ‘imaginary friends’.

Part 2: Descriptive report (990 words + 4 visualisations)

When analysing the datasets, we found some abnormal entries. For example, in the personality ratings, one row displays percentages that appear to be on a scale of 1 to 10 instead of 1 to 100. Additionally, there are duplicates in the 'Last 4 digits of your mobile (0000 to 9999)' column of the table containing personality ratings. It seems that not only are there fewer responses for the personality ratings, but also some rows have a four-digit number that does not correspond with any entry in the table with demographic data. It is assumed that some students only responded to one of the two surveys or used different digits in the second survey, making the merged table an unreliable dataset to work with. Therefore, attention was redirected to the demographics table since it contains the largest number of entries and provides a good opportunity to investigate potential reasons for outliers in an effort to comprehend the apparently arbitrary values in certain rows.

We, hence, conducted a critical examination of outliers to gain a better understanding. This highlights the important role that these anomalies play in revealing underlying patterns, behaviours, or exceptional cases in real-world contexts. The following text demonstrates how outliers serve not only as statistical deviations but also provide valuable insights into unique or atypical phenomena. The aim of this exploration is to reveal the stories behind the data by analysing variables such as age, academic performance, and commuting distances. The results will be visually presented through Power BI, taking advantage of its interactive features to analyse a wide range of data distributions. This will provide a detailed understanding of the diverse student population.

The data analysis reveals the presence of outliers in various segments of the dataset. Demographic information shows that the age distribution mainly centres around individuals in their twenties, which aligns with the expected profile of university students. However, some data points indicate the existence of participants significantly older than the norm, with ages such as 36, 47, and even 58. These deviations may be due to students providing random ages or to the influence of DCU's Age Friendly University initiative. These outliers, although clearly visible on graphical representations, can also be identified through mathematical methods or specific algorithms designed for outlier detection. To systematically identify outliers, the interquartile range (IQR) method can be used. The IQR was used to analyse the dataset as it provides a robust measure against skewed data and extreme values, focusing on the middle 50% to identify outliers effectively. This method ensures a reliable analysis even in diverse data distributions, making it an ideal choice for a comprehensive examination of our dataset.

The calculation for the IQR of the age column would be:

Position of Q1 = $\frac{1}{4} \times (N + 1) = \frac{1}{4} \times (125 + 1) = 31.5$

Therefore, Q1 = 20, as the median of row 31 and 32 = 20.

Position of Q3 = $\frac{3}{4} \times (N + 1) = \frac{3}{4} \times (125 + 1) = 94.5$

Therefore, Q3 = 22, as the median of row 94 and 95 = 22.

(N = total number of data points)

IQR = Q3 – Q1 = 22 – 20 = 2

Lower bound = Q1 – IQR × 1.5 = 20 – 2 × 1.5 = 17

Upper bound = Q3 + IQR × 1.5 = 22 + 2 × 1.5 = 25

The calculation involves subtracting the first quartile (Q1) from the third quartile (Q3), with outliers defined as values more than 1.5 times the interquartile range (IQR) below Q1 or above Q3. In this case, ages below 17 and above 25 are considered exceptional, given an IQR of 2 years.

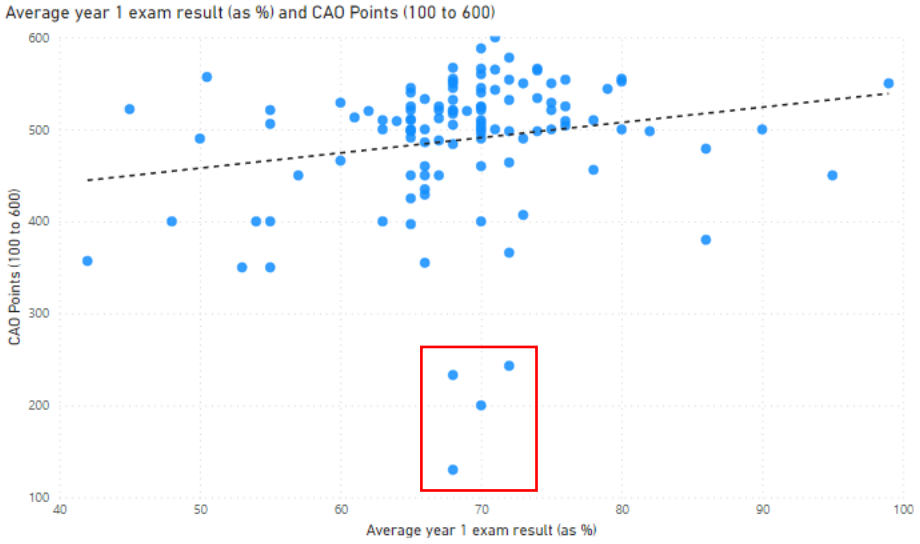
In Python, the code to identify outliers could be the following:

```
In [4]: import pandas as pd

        # Load the data into 'df' DataFrame
        df = pd.read_excel(r"C:\Users\lroes\Documents\DCU\DA Marketing\CA259 Students_first attempt_input.xlsx")

        # Calculate Q1 and Q3
        Q1 = df['Age (in years)'].quantile(0.25)
        Q3 = df['Age (in years)'].quantile(0.75)

        # Calculate IQR
        IQR = Q3 - Q1

        # Define bounds for outliers
        lower_bound = Q1 - 1.5 * IQR
        upper_bound = Q3 + 1.5 * IQR

        # Identify outliers
        outliers = df[(df['Age (in years)'] < lower_bound) | (df['Age (in years)'] > upper_bound)]
        print("Outliers in the 'Age (in years)' column:\n", outliers[['Age (in years)']])
```

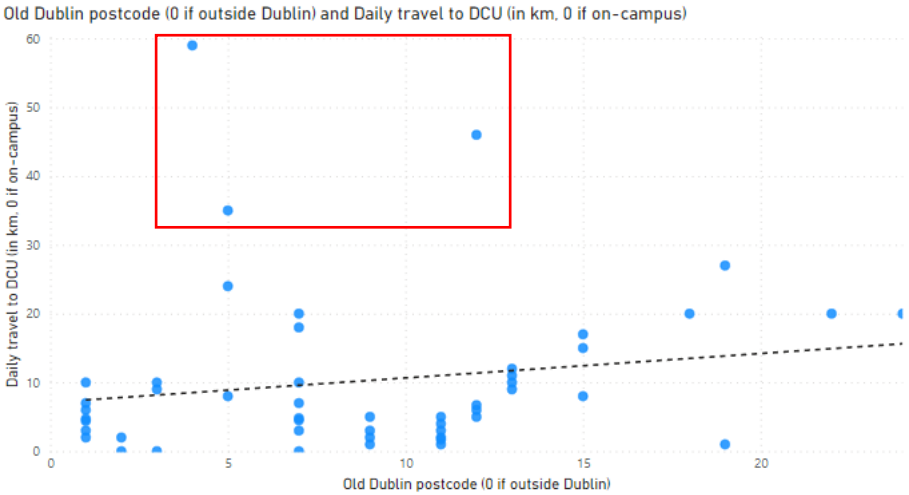
The corresponding output would be the following:

```
Outliers in the 'Age (in years)' column:
      Age (in years)
25                28
29                28
32                36
38                33
60                32
62                26
84                36
87                26
92                33
95                47
103               58
```

Another noteworthy aspect is the correlation between CAO scores and average first year grades. Despite achieving commendable grades ranging between 68 and 72 percent, reflecting second and first-class honours, a minority of students recorded surprisingly low CAO scores, falling below the 250-point threshold, which is typically insufficient for program admission. It is probable that this anomaly occurred because international students chose random numbers, perhaps because they were not familiar with the CAO scoring system.

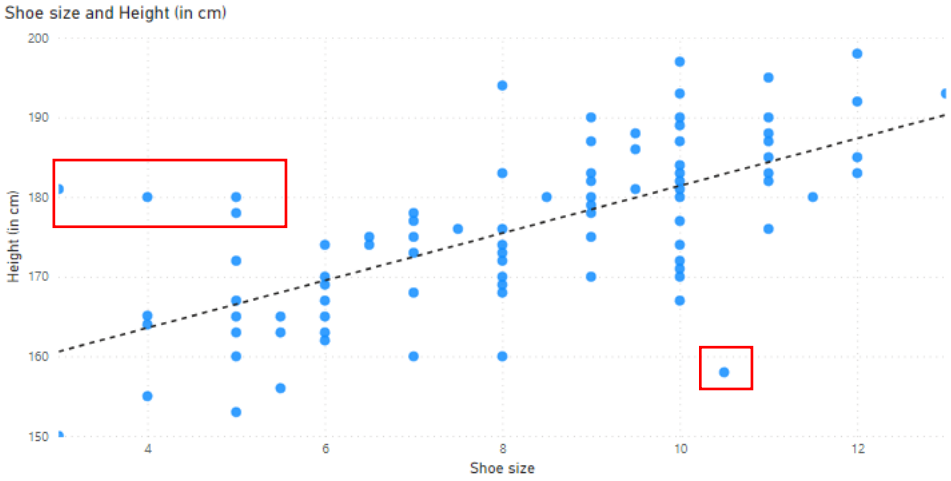


An analysis of commuting distances to DCU campus in relation to students' Dublin postcodes has revealed some peculiar findings. Some students who reside in Dublin but did not indicate '0' to signify living outside Dublin, reported improbable travel distances. For instance, one entry listed Dublin 4 with a commute of 59 km, which is a stark contrast to the actual distance of no more than 10km. These discrepancies raise questions about the accuracy of the entries, indicating potential typographical errors, fabricated figures, or an unconventional way of expressing commute time in terms of distance.



Additionally, the data reveals an interesting pattern regarding the relationship between shoe size and height. While a general trend suggests an increase in shoe size with height, there are numerous exceptions. For example, male students who are 1.80m tall reportedly wear size ten shoes, while a female student who is 1.58m tall has a shoe size of 10.5. This discrepancy

highlights possible confusion over shoe sizing conventions, with some students possibly using UK or US sizing systems, or inaccurately converting from the EU system. Furthermore, the difference between women's and men's shoe sizes could contribute to these anomalies.



In contrast, the correlation between weight and height does not show significant outliers, indicating a proportional increase in weight with height. This consistency suggests a relatively uniform distribution of physical characteristics among the student body, despite considerable variation in heights ranging from 1.58m to 1.98m.

In conclusion, the analysis of outliers in the dataset reveals deviations in age, CAO scores, commuting distances, and shoe size to height ratios. Further investigation into the causes of these anomalies is necessary for accurate data interpretation and decision-making. This highlights the importance of rigorous data validation and the need for awareness of cultural and operational differences that may influence data collection and reporting.