GEA1000: Quantitative Reasoning with Data
Group Project

Tutorial Slot | D39
Group Number | 2

Group Member	**(Matriculation Number)**
Yuan Jing Yi	A0277797Y
Kieran Tan	A0273247B
Eng Cheng Xin	A0288099E
Ho Jian Tao	A0273320R
Edna Ong	A0286583M
Genieve Chin	A0281661A

Part A

Section 1:

1.

- **The aim of the study was** to find out if there is an association between cognitive function and traumatic upper-limb injury through the comparison of the cognitive functions of the patients with and without such injuries.

2.

- **The main independent variable is** traumatic upper-limb injury while **the dependent variable is** cognitive function.
- The upper-limb injury of the participants was recorded in terms of various traumatic characteristics, which includes the time since injury (days), the affected side (left or right), the presence or the absence of nerve injury, hand function, pain level, and sensation.
    - Hand motion was tested with the 4-part Action Research Arm Test (ARAT). Here, the test assesses participants on a scale of 0-3 (where 0 indicates no corresponding action within 60 seconds, where 3 indicates the completion of an action in a normal position within 5 seconds), through grasping (6 items), holding (4 items), pinching (6 items), and coarse motion (3 items).
    - Pain was tested with a self-rated visual analogue scale (0 to 10).
    - Sensation was tested with a Semmes-Weinstein monofilament on the 6-piece foot-kit Touch-Test. 5 levels were used to classify the results into: diminished light touch, diminished protective sensation, loss of protective sensation, and deep pressure sensation only.
- The cognitive function of the participants was measured on the Chinese version of the *Rey Auditory and Verbal Learning Test* (RAVLT), and *Stroop Colour and Word Test* (SCWT).
    - RAVLT is used for the measurement of the aspects of short-term memory in individuals. Here, 12 words are read aloud to the participants who are then instructed to recall the words 3 times. After 5 minutes of an interference test, they are asked to recall the 12 words for a fourth time (delayed free recall). After another 20 minutes (long delayed free recall), they are then needed to recall the words again for the fifth free recall. There are 5 recalls in total and 12 words, with the maximum possible score of 60 points. The score is determined by the number of correct words recalled out of 60, where a higher score indicates a better memory performance.
    - SCWT is used for the evaluation of executive functions, and consists of 3 tasks. Here, the first task is for the participants to read the words that are printed in black. The second task is to recognise the colours of dots, while the third is to recognise the ink colour of the word. There are times when the word does not correspond with the colour of the word in task three. For instance, the word ‘blue’ can be written in red ink, and the participants have to make the effort to filter out the conflict and recognise the colour. The scoring system is both Stroop interference-effect consumed time (time consumed in 3rd task - time consumed in 2nd task) and Stroop interference-effect correct number (correct number in 2nd task - correct number in 3rd task), where a higher score indicates a poorer executive function performance.

3.

- ‘In conclusion, individuals with traumatic upper-limb injury had greater cognitive deficits in short-term memory than uninjured individuals.’ (page 5)

4.

- **The observational group subjects were** the patients from the Guangdong Work Injury Rehabilitation Hospital (Guangzhou, China) who had unilateral traumatic upper-limb injuries. They were admitted to either the outpatient or the inpatient department of the hospital.
    - The patients who were chosen had to fit the inclusion criteria: be able to speak Chinese, be 18 to 64 years of age, be medically stable, have provided consent, and have scored 26 or more points in the Mini-Mental State Examination.
    - The patients with a history of brain or central nervous system disease or trauma (for instance, brain injury, dementia, epilepsy, or stroke), congenital or developmental disease, cognitive impairment (for instance, due to substance addiction), chronic metabolic disease, damage to sight, hearing, smell or taste, pregnancy, history of traumatic events or previous surgery, (for instance, diabetes, hyperlipidemia or hypertension), participation in an interventional trial in the past 6 months, and the identification of psychological problems in the admission screening by the department of psychotherapy were excluded from the study.
- **The control group subjects were** the uninjured individuals who were recruited from the local community. They include the family members of the patients, the nearby community residents, or the factory workers in the city.
    - The people who were chosen had to fit the criteria: be 18 to 64 years of age, be proficient in Chinese, have a normal mental status, and have provided consent.
    - The exclusion criteria were the same as those for the observational group.
- Both the observational groups and the control groups have 104 subjects in their groups, where 70 were males and 34, females.

5.

- **The target population for this study was** individuals with traumatic upper-limb injuries.

Section 2ii: (For an Observational Study)

6. a)

- **Age:** The memory functions of an individual decline as his age increases.
- **Damage to Senses (Hearing, Sight):** The impairment in the senses of an individual might impact his performance in the test because of factors such as poor hearing and sight. For instance, the RAVLT test requires an individual to listen to the words which are read aloud to him. Here, the possession of hearing impairments might impact the result of the test. Furthermore, the loss in the senses of an individual might also lead to other factors, such as a lack of physical activity, depression or social isolation, which might impact their cognitive functions.
- **Medical History:** The past injuries or medical problems of an individual might impact the results of the test. For instance, for an individual with cognitive impairment prior to his traumatic upper-limb injury, the result of the test might not really be because of his upper-limb injury but instead, his past medical history of cognitive impairment.
- **Medical Stability:** The medications which are administered to an individual to aid his recovery might have side effects which might in turn, impact his test results. For instance, some medications have side effects which affect the cognitive and negative functions of an individual.
- **Mental Status and Psychological Well-Being:** The mental conditions and psychological problems of an individual might impact his cognitive and memory functions. For instance, schizophrenia or bipolar disorder might impact the way an individual thinks and his thought processes.

which are obtained from the tests done on the participants might not be due to the main independent variable, but instead the poor cognitive function which might have pre-existed in him, and might in turn, compromise the accuracy of the results of the test.

- There is bound to be progress in the process of the rehabilitation of the participant, which was not acknowledged by the researchers. The participants who were in the middle of rehabilitation might have used more of their cognitive functions in comparison to those who have yet to go through rehabilitation, and might in turn, impact the results of the tests done on the participants.

10.

- The result of the hypothesis test is not reliable to a certain extent because of the absence of some crucial information, such as the pre-injury cognitive functions of the participants, which might impact the results of the test. Without this data, it cannot be determined that upper-limb traumatic injury is the sole reason for the changes in the cognitive functions of an individual.
- The subjects are from the same geographic location (Guangdong). However, the target population, the patients with upper-limb traumatic injuries all across China, might have different changes in cognitive function as a result of their upper-limb traumatic injuries. Hence, the findings of the hypothesis test can only be generalised to a subpopulation of the observational group subjects, the patients from the Guangdong Work Injury Rehabilitation Hospital and were admitted to either the outpatient or inpatient department of the hospital.

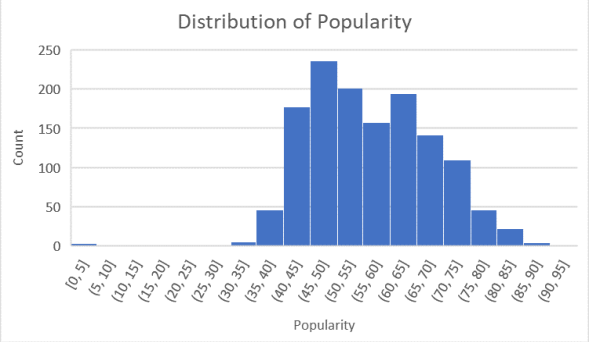
Part B

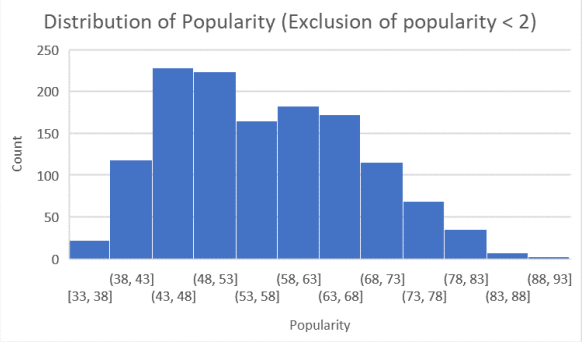
Section 1: Basic Information and Data Transformations

1.

Popularity	
Mean	56.79300074
Median	56
Mode	50
Standard Deviation	11.4247507
25th Percentile	48
75th Percentile	65
Interquartile Range	17
Range	91
Minimum	0
Maximum	91
Count	1343

The number of data points in the data set is 1343. **The categorical variables are** Song, Artists, Explicit, Mode, Key, Release_date, Year, Sub Genre and Genre, while **the numerical variables are** Acousticness, Danceability, Duration_ms, Energy, Instrumentalness, Liveness, Popularity, Speechiness, Tempo, Valence, Loudness and Views.
The calculation of the summary statistics of the data set shows where the average, the median and the mode are in the distribution of Popularity. Here, standard deviation, the measurement of the spread of data points from the mean, is not very wide. The 25th Percentile, the 75th Percentile and the interquartile range also show us the spread of the middle 50% of the data, which is quite compact. It also helps us to determine the most and the least popular song based on the minimum and the maximum value of the variable. In the histogram of the variable, it can also be seen that Mode < Median < Mean, which indicates a potential right skew. After the exclusion of the outliers (where Popularity < 2), the skewness of the histogram becomes more clear.





2.

Tempo Cat	Proportion (%, to two decimal places)
0 ≤ x < 40 : Extremely Slow	0.30
40 ≤ x < 66 : Very Slow	0.67
66 ≤ x < 76 : Slow	2.53
76 ≤ x < 108 : Walking Pace	32.61
108 ≤ x < 120 : Moderate	12.96
120 ≤ x < 168 : Fast	42.00
168 ≤ x < 200 : Very Fast	8.49
x ≥ 200 : Extremely Fast	0.45

Mode is the value which is of the highest frequency in a data set. Therefore, if someone were to say that the mode of Tempo Cat is the category with the highest proportion for this variable, he is correct.

3.

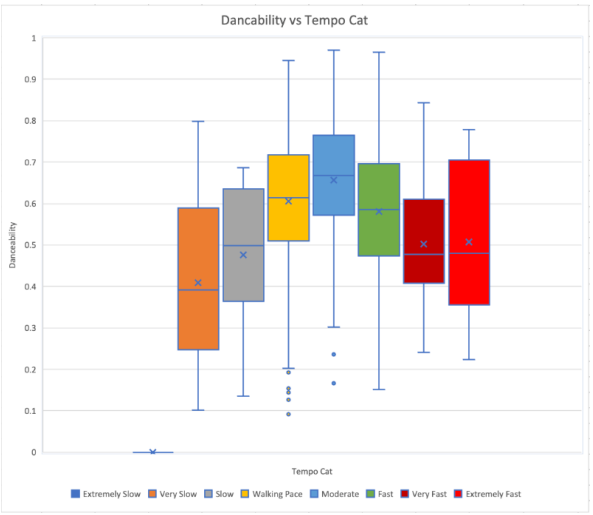
Here, mode is chosen as the additional categorical variable. Based on the Basic Rule on Rates, rate(Happy) = 0.5160 lies in between rate(Happy | 0) = 0.4940 and rate(Happy | 1) = 0.5259.

	Mode					
	0 (Minor)		1 (Major)			
Valence Cat	count	rate	count	rate	Grand Total	rate
Happy	205	0.4940	488	0.5259	693	0.5160
Not Happy	210	0.5060	440	0.4741	650	0.4840
Grand Total	415	1	928	1	1343	1

Section 2: Exploration

4.

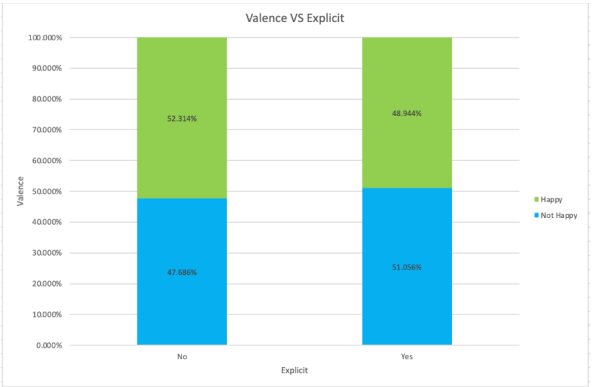
From the boxplot, it can be seen that the songs which have an Extremely Slow tempo have no danceability, while those which have a Moderate tempo have the highest danceability. The songs with the ranges of tempos which deviate from Moderate tempo display lower danceability too. However, there are some ranges of tempo which have very little data, for instance, Extremely Slow (4), Very Slow (9) and Extremely Fast (6). This might, in turn, impact the accuracy of the observation because of the small sample size.



Extremely Slow	Very Slow	Slow	Walking Pace	Moderate	Fast	Very Fast	Extremely Fast
4	9	34	438	174	564	114	6

5.

Explicit is the variable in the data set which comes closest to describing whether a song has vulgarities. Here, Rate(Not Happy | Explicit) = 0.511 is higher than Rate(Happy | Explicit) = 0.489. This shows that there is a positive association between songs which are Not Happy and have Explicit content. However, the extent to which this claim can be answered using the data is limited, because of the small difference in the two rates.



6.

a) Overall, there is a strong negative linear association between Acousticness and Energy, where the r value = -0.713.

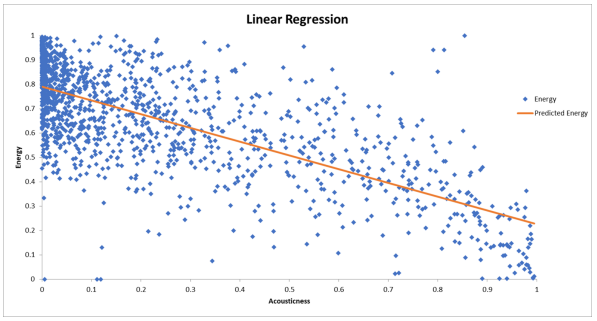
b)

Association	Genre	r-value
• Linear • Negative • Strong	Christian Music	-0.755
	Fusion	-0.779
	Indie	-0.748
	Metal	-0.915
	Norteno	-0.876
	Pop	-0.705
	Rock	-0.804
	Stomp and Holler	-0.774
	Urban Music	-0.724
• Linear • Moderate • Negative	Country	-0.578
	Hip-Hop	-0.358
	Jazz/Blues	-0.677
	Latin American Music	-0.431
	R&B	-0.654
	Rap	-0.351
	Soul	-0.627
• Linear • Negative • Weak	Regional Mexican Music	-0.287
	Reggae	-0.307

The r-values of Regional Mexican Music and of Reggae do not conform to the other two observed associations. Other than those two particular genres, other genres have an r-value that lies near the r-value calculated for the two observed associations. However, in general, all three observed associations are linear and negative; they differ only in terms of the degree of strength, which conforms to the overall negative linear association between acousticness and energy. Therefore, it can be said there is little significant difference between the associations obtained for the different subgroups and the overall association observed.

Here, genres with a data size of three or less were considered to be outliers and were excluded from the table to better streamline the data set.

c)



Based on what has been done for Part a) and Part b), if the genre of the song in question is not in the data set, a scatter plot which depicts the relationship between Energy and Acousticness can be plotted for it, where the independent variable is Acousticness and the dependent variable is Energy. A best fit straight line of the form $y = mx + c$ can then be drawn, by the Method of Least Squares, to visualise the relationship between Energy and Acousticness for it. Here, the Energy of the song can be predicted by interpolation of the best fit line, which is of the equation Energy = - 0.565 Acousticness +

0.7909. This equation can then be used to describe the relationship between *x* (Acousticness) and *y* (Energy) or to make predictions with respect to this relationship. However, the suitability of this equation and/or depends on the specific context and the data it is applied to. It is important to consider the nature of the data and whether a linear model is appropriate for the given situation.

7.

Ecological fallacy is a logical and a statistical error which occurs when conclusions on individuals are drawn with respect to correlations or data which are made at the group level. For instance, in Question 6, there is a strong negative linear association between Acousticness and Energy. However, this relationship does not hold for the genre of Regional Mexican Music, where the r-value is calculated to be -0.287, which signifies that the association is instead, weak. Here, the ecological fallacy is that not all of the genres will have the same strong negative linear association which the overall association implies, and as such, cannot be equated to one another.

Section 3: Inference, Further Analysis and Limitations

8.

The Sample Mean of the average length of Rock songs is 228678.35 ms while the Population Mean of the average length of Rock songs is 232377.17 ms. For Sample Mean < Population Mean, the Alternate Hypothesis will be that the average length of Rock songs is less than 4 minutes and 15 seconds.

```
Single mean test
Data      : Duration_for_hypothesis
Variable  : Duration_ms
Confidence: 0.95
Null hyp. : the mean of Duration_ms = 255000
Alt. hyp. : the mean of Duration_ms is < 255000

        mean   n n_missing         sd        se        me
 228,678.350 183         0 58,696.517 4,338.971 8,561.155

      diff       se t.value p.value  df   0%      95%
 -26321.65 4338.971  -6.066  < .001 182 -Inf 235851.8 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null Hypothesis: The mean length of, *Duration_ms*, Rock songs is 4 minutes and 15 seconds, 255000 ms.

H_0: μ = 255 000 ms.

Alternate Hypothesis: The mean length of, *Duration_ms*, Rock songs is less than 25500 ms.

H_1: μ < 255 000 ms.

At the confidence level of 95%, the p-value is less than 0.001. Hence, at the 5% level of significance, there is sufficient evidence to reject the Null Hypothesis (H_0) in favour of the Alternate Hypothesis (H_1) to conclude that the mean length of, *Duration_ms*, Rock songs is less than 4 minutes and 15 seconds, 255000 ms.

9.

Mediation Analysis was conducted to examine if the Mode of a song is a potential confounder of the association between the Danceability and the Tempo of the song. In the context of the data set, Danceability is listed as a set of Numerical Data, for which was split into 2 ranges of values to categorise the data points into a set of Categorical Data. Here, a data value of more than 0.6 was denoted as ‘Danceable’ while that of less than 0.6 was denoted as ‘Not Danceable’. The same concept was also applied to Tempo, where a data value more than 120 bpm was denoted as ‘Fast’ while that of less than 120 bpm was denoted as ‘Slow’.

	Major	Minor
Danceable (D)	419	232
Not Danceable (ND)	509	193
Column Total	928	415

Here,
Rate (D/Major) = 419/928 = 0.452
Rate (D/Minor) = 232/415 = 0.559
Since Rate (D/Major) < Rate (D/Minor), there is a negative association between Danceability and Mode.

Here,
Rate (F/Major) = 0.503
Rate (F/Minor) = 0.523
Since Rate (F/Major) < rate (F/Minor), there is a negative association between Mode and Tempo.

	Major	Minor
Fast (F)	467	217
Slow (S)	461	198
Column Total	928	415

Since Mode has an association with both the Danceability of and the Tempo of a song, therefore, Mode is seen to be a confounder of the association between the Danceability of and the Tempo of the song.

10.

- One limitation of the use of this data set for analysis arises from the collection of the 125 Sub-Genres of music by the researcher, and the subsequent selection of the first 1000 songs which were listed on Spotify when the Sub-Genre in question was keyed in. Here, there might be more Sub-Genres than the ones which were identified by the researcher and might result in a disparity between the sample population of the 125 of Sub-Genres of music and the target population of all the Sub-Genres of music there is. The results of the analysis of this data set might, therefore, not be generalisable to the whole target population, the different songs in existence.
- One other limitation of the use of this data set for analysis arises from the method which is used for the selection of the songs, which is one of Non-Probability Sampling. This is because of how the first 1000 songs have a definite chance of selection, while those which succeed do not have such a chance at all, which indicates that the different songs of the different Sub-Genres do not have the same chance of selection. to be one of the selected, and those after will not be selected. The results of the analysis of this data set might, therefore, not be generalisable to the whole target population, the different songs in existence.