

THE INTERNATIONAL STATISTICAL LITERACY PROJECT

INTERNATIONAL STATISTICAL POSTER COMPETITION WINNERS

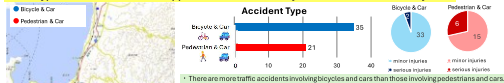
WINNERS: JAPAN

Analysis of traffic accidents involving elementary school students in Matsuyama (2021-2023)

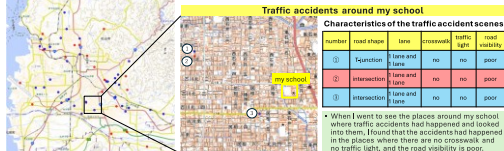
Motivation for research
At my elementary school, there is a traffic safety class for the third graders. In the class, I was curious about the types of the traffic accidents that had happened to the elementary school students in my town, Matsuyama, so I decided to find out more information about them.

Research method
• Collect data about the traffic accidents involving the elementary school students in Matsuyama from the Ehime Prefectural Police Website.
• Go to the places around my school where traffic accidents involving the elementary school students happened and see the characteristics of the traffic accident scenes.

What types of traffic accidents happened to the elementary school students in Matsuyama?



• There are more traffic accidents involving bicycles and cars than those involving pedestrians and cars.
• More people get seriously injured in bicycle and car accidents than in pedestrian and car accidents.

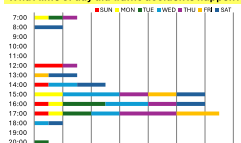


How has the number of traffic accidents changed?



• From 2021 to 2023, the number of the traffic accidents involving the elementary school students went up every year.
• In the past three years from 2021 to 2023, March and June had the most accidents in total, with 8 each.

What time of day did traffic accidents happen?



What was the weather like when traffic accidents happened?



• I thought there were more traffic accidents on cloudy or rainy days, but actually, more traffic accidents happened on sunny days.
• I think the number of accidents goes up on sunny days because it's easier to go out on sunny days than on rainy days.

• When looking at the days of the week, Wednesdays had the most accidents, with 12, and Saturdays had 11 in the past three years.
• When looking at the time of day, traffic accidents often happened between 15:00 and 18:00 on weekdays.
• I think the elementary school students may have been involved in an accident on their way home from school, or on their way from home to their friend's house.

Impressions after completing the survey

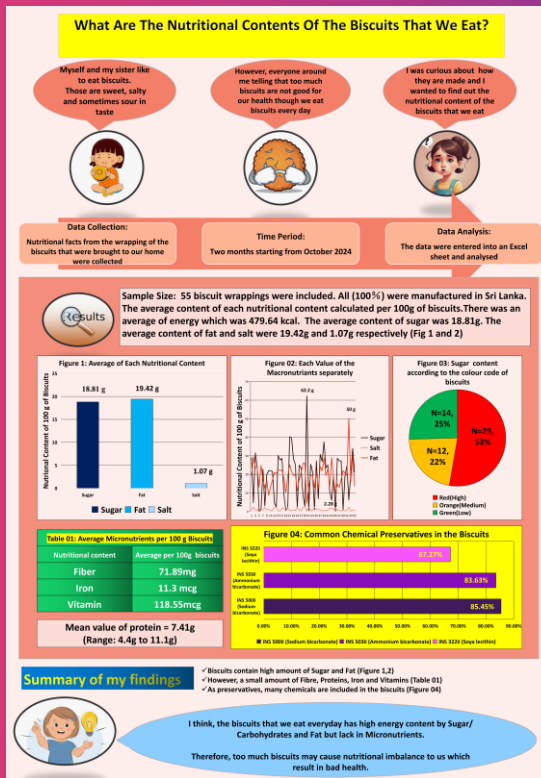
- We could understand the characteristics of the traffic accidents better by showing the details in graphs, tables, and maps.
- Next, I want to research whether the elementary school students wore their helmets or not when they had bicycle accidents.
- I will keep following traffic rules to avoid accidents.

Students: Ayaha Watanabe and Kokomi Watanabe



ELEMENTARY THIRD PRIZE WINNERS: SRI LANKA

Students:
Ashmi
Methusari
Bandara



ELEMENTARY HONOURABLE MENTION: CZECHIA

Students:
Lionel Uhlár

DID YOU KNOW?

THAT EVERY SECOND PRIMARY SCHOOL IN THE CZECH REPUBLIC HAS A PROBLEM WITH TRUANCY? MOREOVER, SO-CALLED EXCUSED TRUANCY IS ON THE RISE, WHERE PARENTS EXCUSE THEIR CHILDREN'S ABSENCES EVEN WHEN THERE IS NO SERIOUS REASON FOR IT. THIS COMES FROM THE ANALYSIS OF ANNUAL REPORTS FROM THE CZECH SCHOOL INSPECTORATE

PROPORTION OF SCHOOLS WITH TRUANCY ISSUES

TYPES OF TRUANCY:

- UNEXCUSED
- EXCUSED
- HIDDEN
- ESCAPE

	2016	2017	2018	2019	2020	2021	2022	2023
MISSED HOURS PER STUDENT	855	933	713	-	-	70	126	117
% OF STUDENTS WITH UNEXCUSED ABSENCES	22	24	24	-	-	-	-	34

IN ANNUAL REPORTS OF THE CZECH SCHOOL INSPECTORATE SOURCE ISI
* THE NUMBER OF DATA IN THE TABLE IS ATTRIBUTED TO THE COVID-19 PANDEMIC

HOW MANY CHILDREN SKIP SCHOOL AND WHY

AVERAGE OCCURRENCE OF BULLYING PER 1,000 STUDENTS BY DISTRICT

NO CASE WAS ADDRESSED HANDLED BETWEEN STUDENTS IN RELATION TO A TEACHER HANDLED OTHER SITUATION

STATEMENTS FROM SCHOOL COUNSELORS ON ADDRESSING BULLYING IN PAST 3 YEARS
FROM THEIR PRIVATE REPORTS BY THE CZECH SCHOOL INSPECTORATE

MANY CHILDREN DO NOT GO TO SCHOOL BECAUSE THEY DO NOT FEEL GOOD THERE. FOR SCHOOL TO BE A PLACE WHERE CHILDREN LOOK FORWARD TO GOING, IT'S IMPORTANT TO OPENLY TALK ABOUT PROBLEMS.

FORMS OF RISKY BEHAVIOUR IN CLASS FROM STUDENT'S PERSPECTIVE

- CYBERBULLYING
- PHYSICAL VIOLENCE
- TRUANCY
- VANDALISM
- INTIMIDATION
- SUBSTANCE USE
- INSULTS
- DISRUPTIVE BEHAVIOUR IN CLASS

LOWER SECONDARY FIRST PRIZE WINNERS: SOUTH KOREA

Students:
Noh Jeongmin

PERCEPTION and BIAS towards AI MUSIC

MOTIVATION

AI is transforming the arts. In music, platforms like DALLE and MID_JOURNEY revolutionize visual arts. Although some fear AI could replace human artists, most experts view it as a creative asset. Yet, persistent biases and challenges in distinguishing AI-generated work from human-made ones highlight the need for further exploration to understand the future role of AI in creative fields.

GOALS

1. Assess individuals' ability to distinguish between AI-generated and human-made music.
2. Determine whether negative evaluations of AI art, especially in music, exist.
3. Examine how these biases are influenced by participants' views on AI's creative capabilities.
4. Analyze how these biases and distinguishing abilities are influenced by individuals' personal characteristics (e.g. gender)

Therefore, this study aims to

PRE-RESEARCH

People rate AI-generated artworks lower than human-created ones across domains like beauty or creativity. (Belluche L., et al. (2023))

Some personal traits correlate with how one reacts to AI artworks.

AI music, especially Classical, is slightly more distinguishable than AI art. (Ferreira P., et al. (2023))

METHODOLOGY

Total Respondent = 146

Survey Process: Participants listened to 4 music (2 made by human, 2 generated by AI), evaluate it on a 7-point likert scale, and guess the composer's identity. They also answered how creative they thought AI was on a 3-point likert scale and who deserves copyright for AI-generated music.

Jan. 13 - 24, 2025

ACCURACY Q1

54.11% of people accurately identified AI-generated music. This suggests that biases didn't happen from the discernible drop in compositional quality for AI-generated music, but rather from people's perception.

While most comparisons were insignificant, Classical listeners were able to discern AI-generated classical music 50.7% more than popular music listeners.

CREATIVITY Q2

AI Music Perception and Creativity Scale

In general, males, Koreans, Classical listeners, and people who believe copyright exist believed AI is creative more than their counterpart.

People who give higher ratings to music they think was composed by an AI also tend to think AI to be more creative.

Who owns the copyright to AI-generated music?

Use of that	69
AI Developer	41
AI Database Copyright Owners	41
No Copyrights at All	12

*Multiple choices were allowed

32 (18.5%) respondents thought no copyright should be given at all.

119 (81.5%) think some copyright should be given.

PERCEPTION Q3

Analysis Method: "Perceived authorship" is the listener's believed composer's identity. "Actual authorship" is the true origin.

When preference score for Q2p 1 & Q2p 2, evaluations are swayed more by who listeners think created the music than by the music's actual quality.

This suggests that the bias stems from the idea of AI authorship making the music less favorable, not because the compositions are inherently less skilful.

People exhibited a significantly greater disparity in ratings based on perceived authorship (left) than on actual authorship (right), indicating that the bias stems from perception rather than from the music's intrinsic quality.

Left mean difference: 1.17
Right mean difference: 0.54

CONCLUSION

1. Just over half (54.11%) of participants correctly identified AI-generated music, suggesting that biases didn't happen from the discernible drop in compositional quality for AI-generated music, but rather from people's perception.
2. A notable perception gap emerged: music believed to be AI-composed received lower ratings—disparity greater than that observed based on actual authorship. This indicates that the bias stems from perceptions of non-human authorship, not from the inherent quality of the music.
3. Skepticism regarding AI's creative abilities grew, leading to less and unreserved issue surrounding copyright and ownership.

SO WHAT?

Unreserved issues around AI-generated music and unclear legal boundaries to public occurrence of AI art. Addressing these issues could lead to a clearer understanding of AI's role in the creative process.

IMPROVEMENTS

1. Broaden the sample size and ensure demographic diversity.
2. Refine the experiment design by including a control group, where authorship is revealed only after initial ratings.
3. Incorporate more qualitative feedback through open-ended questions.



LOWER SECONDARY SECOND PRIZE WINNERS: SINGAPORE

Students:
Lai Wei Qii,
Jocelyn Lau Yu Xuan,
Goh Fay Jinn

HOW ACCURATE ARE THE WEATHER FORECAST WEBSITES IN SINGAPORE?

Research question & Motivation
Our research focused on investigating the accuracy of weather forecast websites in Singapore, namely:
AccuWeather, BBC Weather, MSE & NEA forecast

AccuWeather
We were inspired to focus on this due to our real life experiences, where we checked the weather forecast beforehand, but since we watched our destination, the weather was not the same as predicted. Thus we wanted to see which commonly used weather forecast is the most accurate, to allow the transportation and activities planned by Singaporeans to be able to get going straight in the rain or suffering from heat stroke.

Introduction
With climate change seeming to make weather patterns more erratic, such as recent flash floods that have been plaguing the nation, another statistic has become increasingly essential in order to carry out everyday tasks. Thus finding the most reliable weather forecast that the public can rely on is of utmost importance.

Purpose and hypothesis
We want to find the most reliable commonly used weather forecast in Singapore. Weather forecasts are used daily by everyone and play an extremely crucial role in our lives. It does not merely help us decide what clothes to do and what items to bring, it can save lives by giving early warnings of storms, hurricanes and other natural disasters and determine the amount of forecast errors social media. In Singapore's context, it can help us to predict flash floods, reduce traffic congestion to prevent accidents. Our hypothesis is that between AccuWeather, BBC Weather, MSE and NEA forecasts, **NEA forecast is the most accurate for both forecasting of temperature and rain rate and that local weather forecast websites are more accurate than foreign weather websites for both.** We also hypothesized that **AccuWeather and BBC weather are more accurate, with MSE being the least accurate weather forecast in Singapore and NEA forecast is managed by the local authorities.**

Methodology
The method we used for our hypothesis was to collect the daily forecasted temperature and presence of rain from AccuWeather, BBC Weather, MSE and NEA forecast before comparing the collected data to the raw data recorded by NEA. To collect the required data, we used each website to collect the temperature and presence of rain in 4 neighbouring areas (Singapore island) on the website at 12 PM daily, which includes today and 1 website for the largest amount of data that we require simultaneously. We collected the data over a period of 4 months.

Temperature
To find the accuracy of the weather forecasts, we compared the predicted minimum and maximum value to the actual values recorded by NEA using their Absolute Percentage Error (APE), Weighted Absolute Percentage Error (WAPE) and Mean Squared Error (MSE) which they do for each weather forecast. We used a 1 to 5 metric to measure the forecasts that each group. Group WAPE measures the accuracy of weather forecasts between forecasted and actual values in a percentage. Group MSE is similar to WAPE but with WAPE distribution weights based on magnitude of actual value, normalizing the error relative to overall size of data while MSE gives more weights to larger errors, making the errors more prominent.

Rain
Let n = number of forecasted observations, A = actual value and F = forecasted value.
Mean Absolute Percentage Error (MAPE) = $\frac{1}{n} \sum_{i=1}^n \frac{|A_i - F_i|}{A_i} \times 100\%$
Weighted Absolute Percentage Error (WAPE) = $\frac{1}{n} \sum_{i=1}^n \frac{|A_i - F_i|}{A_i + F_i} \times 100\%$
Mean Squared Error (MSE) = $\frac{1}{n} \sum_{i=1}^n (A_i - F_i)^2$

Data Analysis

Figure 1: Bar chart showing MAPE for different weather forecast websites. AccuWeather: 15.1%, BBC Weather: 17.7%, MSE: 16.2%, NEA: 10.5%.

Figure 2: Bar chart comparing MAPE for temperature and rain rate forecasts. Temperature: AccuWeather (15.1%), BBC (17.7%), MSE (16.2%), NEA (10.5%). Rain: AccuWeather (15.1%), BBC (17.7%), MSE (16.2%), NEA (10.5%).

Figure 3: Line graphs showing temperature and rain rate forecasts over time for different websites.

Conclusion
This proves that our hypothesis is correct. Although NEA forecast is most accurate for forecasting of presence of rain just like our hypothesis, however unlike what we hypothesized, AccuWeather is most accurate for the temperature forecasting. This means that the commonly used weather websites are generally better at predicting if it would rain, while foreign weather websites are generally better at predicting the temperature. This may be due to several reasons, local weather forecasting websites may be able to better predict if it would rain due to their better understanding of local weather patterns which is less predictable than temperature, while foreign or international weather forecasting websites might be better at predicting temperature as they have better access to data of air circulation.

Future Applications
This could be used to help Singaporeans to be better prepared to deal with the uncertainty of different weather forecasts during occurrence of different weather phenomena. Likewise, a larger variety of weather forecasting websites could be included so that we can use which foreign weather forecasting websites is the most accurate and potentially use various kinds. For example, if weather forecasting app used to inform us are generally more accurate.

Limitations
One limitation for the lack of data, resulting in a insufficient sample size that might cause the analysis to be less reliable due to the emergence of outliers, such as different weather patterns due to climate change. When sudden extreme weather events occur, weather forecasts might be unable to predict it causing a increase in the error value for the different weather forecasts. Moreover, with only 4 websites used, we are unable to test the accuracy of the weather forecasts during the occurrence of different weather phenomena which occur at different parts of the year.



LOWER SECONDARY THIRD PRIZE WINNERS: CANADA

Students:
Brendan Cai

BACKGROUND AND OBJECTIVE OF THE STUDY

I began to get interested in my current high school in September 2024. I've noticed that [transportation](#) is a big issue. The school bus departs at 8:00 AM, and it's hard to be there by 8:00 AM on some days. My parents and I have talked about this, but they don't seem to have a solution. This has led me to wonder [how to solve this problem](#). I've decided to do a project on this.

I've been thinking about statistics and how it can help people solve better decisions every day. That's got me wondering [how to solve this problem](#). I've decided to do a project on this.

With these questions in mind, I decided to learn more about statistics and collect data to solve the problem for me and my parents.

CAN STATISTICAL ANALYSIS HELP ME ARRIVE AT SCHOOL ON TIME?

ANALYSIS OF DATA COLLECTED IN PHASE ONE

Travel time data was collected over 10 phases (Oct. 10 - Oct. 20, 2024). During this period, I arrived after the targeted arrival time of 8:00 AM on 10 days, and after the best time of 8:00 AM on 10 days.

See plots and summary statistics of daily late travel times from my home to school (Start, waiting times at the bus stop, and driving times) along the route (Phases 1, 2, 3, 4, 5, and 6) during times including waiting times at red lights. Driving are presented. Daily total travel times include a large variability, generally due to fluctuations in Red 2 and Driving. An outlier indicated in my data set was noted. This outlier coincides with the outlier in Driving, and the most extreme outlier in Red 2. The mean of daily total travel times is slightly greater than the median, indicating that frequent travel times tend to be significantly prolonged.

The Start times show considerable variability, with the variability being less affected by my parents' unpredictable morning schedules. The Arrival times exhibit slightly greater variability in terms of SD and IQR than the Start times. This was due to the combined effect of the variability in Start times and daily total travel times.

STRATEGY TO AVOID LATE ARRIVALS

To arrive at school before 8:00 AM, I set a target Start time using the following formula, based on the data collected in phase one:

Target Start Time = 8:00 AM - (Total Travel Time - 7:30 AM)

The 50th percentile of daily total travel times is used in the formula instead of the median or mean to account for potential unexpectedly long travel times due to changing conditions. The goal of start times is to ensure that no traffic jams delays caused by uncertainty in the Start times.

OVERVIEW OF THE APPROACH AND DATA COLLECTION

Daily travel times from my home to school were collected in two phases. The first phase was from October 10 to December 20, 2024. On each school day during this period, I recorded the time I left home (Start Time) and the time I arrived at school (Arrival Time). Additionally, I tracked the waiting times at red lights along my route to school.

After phase one was completed, I analyzed the collected data during the Christmas break using basic statistics and developed a strategy according to the analysis to adjust my Start time so that I could arrive at school on time.

Since January 6, 2025, my parents and I have followed the strategy. Daily travel times were recorded from January 6 to February 4, 2025, marking the second phase of the study. The data collected during this phase were used to evaluate the effectiveness of my strategy.

MAIN STATISTICAL CONCEPTS USED

- Outliers:** A data point that is significantly different from the other data points. In this case, the outlier is a data point that is significantly higher than the rest of the data. The 50th percentile is also called the median.
- Interquartile Range (IQR):** The difference between the 75th percentile and the 25th percentile. The IQR reflects the variability of the dataset.
- Boxplot:** A figure displaying some summary statistics. It includes:
 - A box that represents the IQR of the data, which opens vertically from the 25th percentile (Q1) to the 75th percentile (Q3).
 - A horizontal line inside the box that represents the median of the data.
 - Two vertical lines, known as whiskers, that extend from the bottom and top edges of the box to the smallest and largest data points within 1.5 times the IQR from Q1 and Q3. Data points outside of this range are considered outliers—observations that deviate significantly from the rest of the data.
 - Mean:** The average value of the dataset.
 - Standard deviation (SD):** A quantity that measures the variability of the data relative to its mean.

CONCLUSIONS FROM PHASE ONE

- A primary cause for my late arrivals in phase one is the late Start times. The mean Start time was 7:52 AM, and the mean daily total travel time was 8:12 AM, resulting in an average tardiness of 18:00 AM.
- The large variability in both the Start times and travel times, which were driven by uncertain factors, led to unpredictable Arrival times.

EVALUATING THE STRATEGY USING NEW DATA

Following my strategy, my parents and I have been aiming to leave home for school at 7:30 AM since January 7, 2025. Phase two (Jan. 6 to Feb. 4, 2025) of the study. The same five variables were recorded over 10 days. I arrived after 8:00 AM on 10 days, and after 8:00 AM on 10 days.

During phase two, there was one outlier Start time, which was caused by an unexpected event at home. This outlier led to one of my two late arrivals.

The mean Start time was 7:32 AM, which is later than the targeted Start time. However, this isn't a problem because whether I Start time was accounted for my strategy. The SD of Start times was 4:30 minutes, which is smaller than that observed in phase one.

A few outliers are observed in daily total travel times. Red 2 and Driving, most of which was caused by a temporary construction that closed down the traffic near red light 2. This disruption caused my other late arrival in phase two.

Overall, the role of my late arrivals in phase two was significantly reduced compared to phase one. The late arrivals in phase two were due to uncontrollable factors. Given my current situation, further reducing the average Start time seems difficult. However, with effort, the variance in Start times may be reduced so that the Arrival times can be no longer considered in the future.

FINAL CONCLUSION

- By collecting and analyzing data using statistics in phase one, I came up with a strategy to reduce late arrivals. This strategy is proven to be effective based on the data collected in phase two.
- I learned some useful statistics and data properties, and I was convinced that more statistics can help us solve problems in daily lives.
- I learned that plans will be disrupted by uncertain events. By incorporating uncertainty into the planning process, we can enhance the resilience of plans against potential disruptions.



LOWER SECONDARY HONOURABLE MENTION: JAPAN

Students:
Kenichi Suzuki

What shorten healthy life expectancy ?

1 Reality and Problems

Q1 The life expectancy in Japan is about 10 years longer than the world average for both male and female. Japan's healthy life expectancy for both male and female is more than 10 years longer than the world average.

Q2 (Concept) Life expectancy at birth — The expected number of years a 0-year-old child will live.

Q3 (Concept) Healthy life expectancy (HLE) as both — A period during which you can live an independent life. A period when there are no restrictions on the conducting period. — The difference between average life expectancy and healthy life expectancy.

Q4 (Concept) Usually, there are no unhealthy periods. But in Japan, there are periods in which life is restricted for about 10 years for male and more than 12 years for female. What shorten healthy life expectancy ?

2 Hypothesis and Plan

Q1 (Hypothesis) Prefectures with short healthy life expectancies may have diseases with a large number of patients.

Q2 (Hypothesis) Prefectures with long healthy life expectancies have fewer patients with diseases.

Q3 (Plan) I thought of three steps to confirm my hypothesis.

- Step 1: Examine healthy life expectancy by prefecture.
- Step 2: Find out which diseases have the highest number of patients in Japan by gender.
- Step 3: Check the number of patients by prefecture and gender, and compare the relationship with HLE.

Q4 (Plan) There are only a few types of statistical data for each prefecture and gender.

- Disease groupings are based on the Ministry of Health, Labor and Welfare's classification.
- There were many names of diseases that I didn't know.

3 Analysis --- Step 1

Q1 Healthy life expectancy relationship between male and female (2022)

Q2 Prefectures with long healthy life expectancies for both men and women are concentrated in the Chubu region.

Q3 I looked into healthy life expectancy by gender for each prefecture. Of the 47 prefectures, the prefectures that were in the top 10 for both men and women are colored green. Prefectures that were in the top 10 for both men and women are colored red. A positive correlation was found between men's and women's healthy life expectancy.

4 Analysis --- Step 2

Q1 Major diseases with a total of over 1 million patients by gender (2022)

Disease name	Male (100 people)	Disease name	Female (100 people)
Hypertensive disease	746	Benign renal disease	865
Ischemic heart disease	671	Upper respiratory tract disease	844
Heart disease	585	Arthritis	844
Stroke disease	517	Long-term respiratory disease	780
Diabetes	487	Heart disease	740
Alzheimer's disease	487	Heart disease	640
Chronic liver disease	317	Chronic liver disease	318
Senile macular degeneration	217	Senile macular degeneration	218
		Senile macular degeneration	185

Q2 The diseases are listed in order of the number of male and female patients in Japan. The following four types of diseases with more than one million patients. There are several diseases whose distribution both male and female.

Q3 (Reference) The population of Japan is
 *Male 60,462,000 people *Female 43,820,000 people
 *Total 124,282,000 people
 Source: Population estimate (as of October 1, 2022)
 Statistics Bureau of Japan

5 Analysis --- Step 3

Q1 (Hypothesis) The prevalence of diseases listed in Step 2 is related to healthy life expectancy.

Q2 (Plan) I investigated the prevalence of the diseases listed in Step 2 by gender and by prefecture. Next, I investigated the relationship between prevalence and Step 1 healthy life expectancy. Glaxosone was the only disease that showed a negative correlation in male. The only female case was glaxosone.

6 Conclusion

Q1 (Understanding analysis results) In Step 2, I checked diseases in male, which showed a negative correlation.

Q2 In female, I checked dyslipidemia, which showed a weak positive correlation. Most diseases had no relation to healthy life expectancy.

Q3 (Judgement of hypothesis) I found a disease that refutes hypothesis 1. Both male and female had Glaxosone.

Q4 I also found diseases that did not satisfy hypothesis 1. That is, Glaxosone in female.

Q5 The results confirmed that both healthy life expectancy based solely on disease prevalence.

Q6 (Conclusion) Healthy life expectancy is self-reported, so the data from unhealthy periods accurate 7 in the future, we will need something that can determine when people are unhealthy.

Q7 If you are worried about your health, it is important to get to the hospital and check your health. If you start early, the condition may not become life-threatening. As a result, period of unhealthy health will be shorter.



LOWER SECONDARY HONOURABLE MENTION: ITALY

Students:
 Marco Martini,
 Amantia Xhafa,
 Aurora Donanzan,
 Nicolas Fattore,
 Raffaele Grandesso

THE RENEWABLE ENERGY USED IN ROSSANO VENETO

Survey conducted on the population of Rossano Veneto

INTRODUCTION
 We are lower secondary school students from the town of Rossano Veneto, and we conducted a survey on the renewable energy sources used in the homes of our area. We decided to carry out this investigation to raise awareness among people and make their homes more sustainable.

RENEWABLE ENERGY is a type of energy that is produced thanks to **renewable sources** and allows to not damage the environment.
The main ones are:

- Solar energy
- Wind energy
- Geothermal energy
- Biomass energy
- Marine energy
- Hydropower

CHOICE OF THE INVESTIGATION
The objective of this survey is to understand the number of people in Rossano Veneto who use **renewable resources at home**. This allowed us to understand the **consequences of our actions** and what we can improve. We hope that this work will bring more **awareness** to people about their homes and habits.

SURVEY MODE

- Creation of a Google form, which we sent within our school and asked students, their parents, and relatives living in Rossano to fill out, to gather more data.
- Analysis of the collected data.
- Creation of the statistical graphs, in which we included information, which we included information, which we included information.

2030 AGENDA GOAL

To ensure access to affordable, reliable, sustainable and modern energy systems for all
 Goal 7 was created out of concern for rising energy costs and to enhance energy security.

CONCLUSIONS AND THANKS

We hope this work will be used **among residents**, so that they about these sustainable practices and help them **more** make their **contribution** to their own homes to become more **sustainable**. We thank you for giving us the opportunity to participate in this contest, and for making it possible for us to **discover** the **other** students by working together in a **simple** and **easy** way to have completed the investigation and **thank** you for your attention.

BIBLIOGRAPHY

- STAT
- Agenda 2030
- Agri-Campus

Main sources used to heat the house

Electricity	41%
Autonomous gas	34%
Other	24%
Biomass	1%

Interventions done to improve energy efficiency

Thermal insulation	40%
Installation of solar panels	30%
Insulation of their pipes	20%
Energy saving condenser	10%
Exchanging boiler	10%
No intervention	10%
Light bulb	10%

Type of energy that powers household appliances

Electricity	46.9%
Autonomous gas	33.3%
Other	13.5%
Biomass	6.3%

What are solar panels used for

Hot water	48.5%
Electricity	32.3%
Autonomous gas	14.5%
Other	3.7%



LOWER SECONDARY HONOURABLE MENTION: IRELAND

Students:
Eli John Kiernan

BOWLED OVER: A Statistical Analysis of Bowling in Cricket Analysing both the Male and Female Aspects of the Game Using Live Data

Abstract

I carried out a statistical analysis of bowling style and performance metrics in cricket. I concentrated on the limited overs Twenty-Twenty (T20) cricket league only. The format is consistent across the 3 leagues in relation to duration and length of the matches. T20 is the most popular form of cricket because of its quick pace and high stakes require. The use of different bowlers and the need to take wickets makes it a particularly important in the limited overs match format. I investigated six different questions across both male and female leagues (listed below). These questions looked at different aspects within bowling in cricket. These questions are important to understand the impact of bowling style on performance metrics and the findings can be used to inform managerial decision making when choosing which bowler to use in certain circumstances. I was excited to analyse these questions because they are of significant importance to me in my sporting career. I currently play club cricket for Malahide, represent for Fingis and provincial for Leinster. I hope to play for Ireland in the future.

Experimental Methods

I choose T20 cricket leagues because they are limited to 20 overs per innings which allows the leagues to be compared without different match format and makes engaging data. The male cricket leagues I used were the Inter-Provincial League (IPL), the Big Bash League (BBL) and the Women's Big Bash League (WBBL). The female cricket leagues I used were the Women's Big Bash League (WBBL) and the Women's World Cup (WWC).

My research questions:

- Does height affect the bowling style of a bowler?
- Does bowling style affect the number of wickets taken by a bowler?
- Does bowling style affect the number of maidens bowled by a bowler?
- Does bowling style affect the strike rate of a bowler?
- Does bowling style affect the bowling run rate economy (BRR) of a bowler?
- Is there a correlation between the number of wickets taken by a bowler and the number of runs conceded to the batting team by a bowler? Does the correlation vary with bowling style?

Bowling performance metrics were downloaded into excel. Height and bowling style were added manually. This task was a lot more challenging for the female aspect of the game.

The following statistical calculations were completed in excel: Average, standard deviation (STDDEV), STDEV.P, linear regression and Pearson's correlation coefficient. All data is available in the project booklet.

Comparative Data Analysis

Most of the findings were the same between the male and the female leagues such as bowling run rate economy and the correlation between the amount of wickets taken and the amount of runs conceded to the batting team. There were a number of statistical findings different to that of the male leagues such as Pace bowlers in the male leagues took more wickets than that of spin bowlers. In the female leagues it was the opposite in that the WBL and the WWC, with spin bowlers taking more wickets than Pace bowlers.

In the male leagues pace bowlers had lower strike rates than their spin bowling counterparts. The over the strike rate the baller, in both female leagues pace bowlers had higher strike rates, but in their defence, the difference was marginal.

Results and Data Analysis

1. Does height affect the bowling style of a bowler?

Pace bowlers are taller on average than spin bowlers across all 3 leagues. In the female leagues the difference was much smaller when compared to the male leagues. This is not surprising as pace bowlers rely on the speed and swing of the ball which is generated by the height of the bowler's front leg and speed upon release. The taller the bowler is, the higher the release point. A longer stride increases swing speed and the ball is released at a faster pace.

Spin bowlers were taller on average than IPL bowlers. The BBL is the Australian national league and the IPL is the Indian national league. The average height of an Australian man is 175cm and the average height of an Indian man is 165cm.

2. Does bowling style affect the number of wickets taken by a bowler?

Male pace bowlers took more wickets than spin bowlers across all 3 leagues. Female spin bowlers took more wickets than pace bowlers, about the difference was marginal.

3. Does bowling style affect the number of maidens bowled by a bowler?

On average, pace bowlers bowled more maidens than spin bowlers across all 3 leagues. The difference was most notable in the IPL and WBL. The IPL and WBL are considered the highest standard cricket leagues for their gender and the world and attract the best players.

4. Does bowling style affect the strike rate of a bowler?

Pace bowlers had lower average strike rates than spin bowlers across all 3 male leagues. The IPL and BBL average strike rates are consistently lower than the WWC2023 average strike rates. Possible reasons for the higher strike rates in the WWC2023 is the varying skill levels between teams in the 2023 World Cup, whereas in both the IPL and the BBL, the teams would be of similar skill standard. In the female leagues spin bowlers had a lower strike rate than pace bowlers. Reduced height is likely a contributing factor to the speed generated by female pace bowlers.

5. Does bowling style affect the bowling run rate economy of a bowler?

Spin bowlers had a slightly lower bowling run rate economy across all 3 leagues. There is a greater difference in the male leagues.

6. Is there a correlation between the number of wickets taken by a bowler and the number of runs conceded to the batting team by a bowler? Does this correlation vary with bowling style?

Using 2/1 scatterplots, linear regression and Pearson's correlation coefficient (R) there was a strong correlation (-0.5) between the number of wickets taken by the bowler versus the runs scored by the batting team against that bowler in both the IPL (R=-0.505), BBL (R=-0.5445) and WBL (R=-0.5412) when all bowling styles were included in the analysis. However, in the World Cup (WC2023) there was only moderate correlation (-0.4) with R=-0.5059 in the WWC2023 and R=-0.5417 in the WBL.

Implications for cricket

These statistical findings can be used to influence managerial decision making to ensure strategic use of bowlers. The choice of bowler can be tailored to the situation in a match. In T20 cricket it is important for opening bowlers to take more wickets quickly but that the opening bowler takes fewer wickets overall. If the bowler is taking more runs, the risk is that it will be difficult to hit the stumps in the early part of the innings which therefore means the higher chance of boundaries (4 or 6 runs). However, the research is concerning that bowlers bowled before they have a chance to make a large impact on the score. Male pace bowlers should be used at the start of a T20 game to dismiss top order batters quickly. However, the spin bowlers might be beneficial to spin in female T20. In general, spin bowlers are better to use in a situation where you want to conserve run rate and are willing to sacrifice being wicket taking. This could be most useful when the batting side (opposing team) is chasing down a low total when batting second. Height is a key element of bowling to allow you can generate more bounce and spin upon delivery of the ball.

Conclusions

I believe that all of these statistics are of vital importance in the growing industry, not just with managerial decision making but also for a person that is new to the game and who is looking for information on what type of bowler to become and what they want to achieve in cricket. I believe that all of these statistics can be used across the range of cricket from club to international level.

The statistical analysis of bowling metrics in cricket offers valuable insights into performance trends and influencing factors. Gender-specific studies underscore the need for tailored training approaches, while advanced statistical tools enable deeper understanding and prediction of success factors.

References

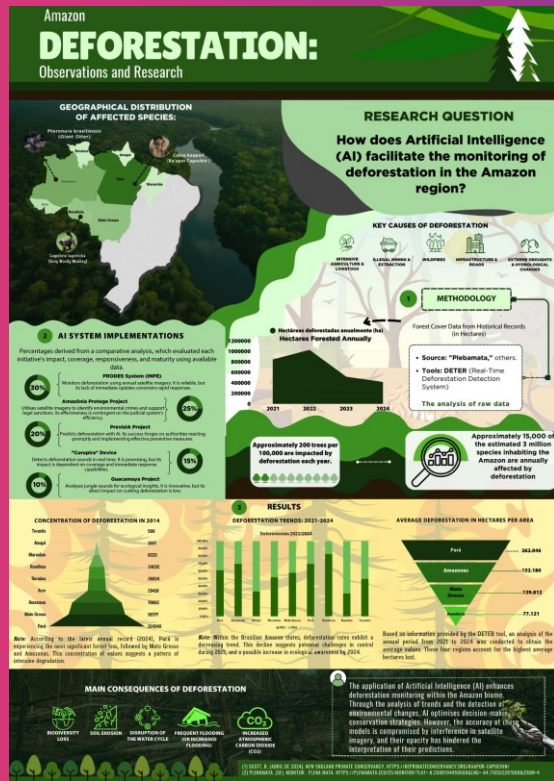
- Shahmoradian, D. and Choi, D. (2024). "A Form-specific Bowling Performance Measure of Cricket." *International Journal of Statistical Science*, Vol. 2(2) Spring, December, 2024, 10-122
- Chen, P., Wang, C. and Wang, S. (2022). "A new uniform and reliable deep player performance index for player evaluation in T20 Cricket." *Decision Analysis Journal*, Vol. 2, 4, 10, 2022
- Chen, P., Wang, C. and Wang, S. (2021). "An ICC2020 approach for evaluating bowler performance in IPL." *Journal of Emerging Trends in Computing and Information Science*, Vol. 16, 10, 1028-1031
- Farooq, P.J., Lewis, S.L., Washington, P.J. and King, S.A. (2019). "Comparison of biomechanical characteristics between male and female fast bowlers." *Journal of Sports Sciences*, 37(10), 858-870
- Thompson, B. G. and Swanson, S. M. (2012). "An interdisciplinary application of principal components to cricket data." *Journal of Statistical Education*, Vol. 21, No. 2
- van den Tillaer, R. L. and Sherrin, C. (2005). "Effect of body size and gender on overhead throwing performance." *Scandinavian Journal of Applied Physiology*, 81, 474-478
- Washington, P.J., King, S.A. & Reeves, C.A. (2014). "Relationship between fast bowling technique and ball release speed in cricket." *Journal of Applied Biomechanics*, 29, 15-24
- http://www.espn.com/cricket



UPPER SECONDARY FIRST PRIZE

WINNERS: ECUADOR

Students:
 Andreina Natalia Merchan Sánchez,
 Christopher Andrew Cobos Baque,
 Saul Eduardo Choez Tapia,
 Kalil Roberto Mera Reyes,
 Mya Katuska Plaza Plúas



WINNERS: **BOLIVIA**

Students:
 Maya Cabrera Tarquino,
 Wara Cabrera Tarquino,
 Isabella Arias Stelzer,



UPPER SECONDARY

HONOURABLE MENTION: JAPAN

Students:
Kento Usami, Yuma Oya, Atsushi Kato, Koki Hori

Save a local town in 2040 by Bus+Mobile Catering

Nagoya University Affiliated Upper and Lower Secondary Schools

0 Our Target Municipality and the Current Situation

Utilize a community bus and retail refuges, helping the elderly's daily shopping in a aging society, with a bus equipped with mobile catering.

Problem 1 Crisis of Community Bus

Problem 2 Will Retail Stores Disappear?

Problem 3 Elderly Shopping Refuges

As the burden of cost for the bus on the town has increased, there are on the verge of extinction.

The number of retail stores in Minamichita was high in the past, but has decreased significantly.

Elderly people are having difficulty shopping, which can have a negative impact on their health.

It is proposed to increase new users and new income source is necessary.

It is proposed to increase sales opportunities of retail stores is necessary.

It is proposed to save elderly shopping refugees and help their health is necessary.

4 Our Suggestion -Urlikko Bus-

Based on analysis so far, Minamichita faces three challenges:

- 1 Public Transportation Maintenance
- 2 Retail Stores Decline
- 3 Elderly Retail Access

To address these challenges, we proposed "Urlikko Bus" - a modified version of "Urlikko Bus" designed specifically for mobile catering services.

Solve these three problems with the "Urlikko Bus."

5 Places to Operate Urlikko Bus Service

6-1 Verification 1 -A Precedent Case-

According to a precedent case, the majority of users of a bus with mobile catering are the elderly, and food sales are high.

6-2 Verification 2 -A Plan in Detail-

Offering daytime sales to the elderly in a way without affecting the current bus schedule and reserving spaces for retail stores for sale promotion during the daytime is necessary, so that bus seats can be reserved and be small shares of "sales space" for the purpose of the elderly.

1 New users and stable income for Urlikko bus 2 Increased sales opportunity for retail stores 3 Helping elderly shopping refugees

The proposal will provide a place for local residents to relax and a chance for the elderly to go out and chat with each other. The whole town will be revitalized through "Urlikko Bus".

Reference



UPPER SECONDARY HONOURABLE MENTION: SPAIN

Students:
Lucía García López,
Carmen Menéndez Osorio

Do we have the Text Neck Syndrome?

Symptoms: For your head posture (PAP), self-book, neck and back pain.

Objective #1: To examine the effect of the time the mobile phone is used (X) on the nature position of the head (Y) in teenagers.

Objective #2: To examine the effect of the observed posture (X) on the neck pain (Y) in teenagers.

Methodology: Descriptive and inferential statistics. Surveys and questionnaires.

Data Collection: Online questionnaire, Smartphone use, Record sheet, Data entry.

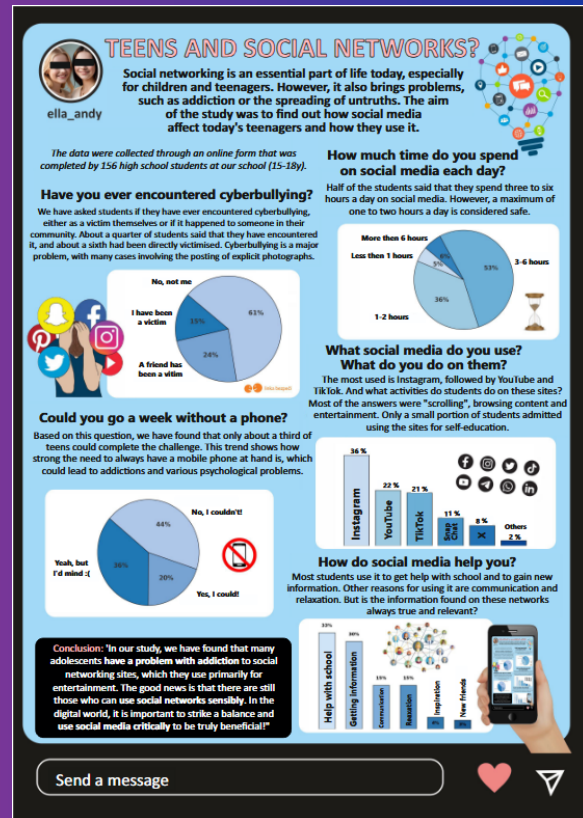
Data Analysis: Descriptive statistics, Inferential statistics (t-test, ANOVA, Chi-square).

Conclusions: Our findings suggest that young people of our age haven't developed Text Neck Syndrome yet. With 95% confidence, we can assert that: No difference of the nature head position existed to mobile phone use was found. Including prior body conditions, there is no evidence to suggest that neck pain is linked to mobile phone usage time of posture.



UPPER SECONDARY HONOURABLE MENTION: CZECHIA

Students:
Eliška Koderová, Anna
Menoušková



UNIVERSITY SECOND PRIZE WINNERS: URUGUAY

Students:
Facundo Morini

Logistic Regression in Forensic Odontology Sex determination through dental measurements

In the field of forensic odontology teeth are an excellent research material. Furthermore teeth are almost an indestructible material. The information provided by their shapes and sizes allow us to determine some characteristics of an individual such as sex or age. This research seeks to evaluate the efficiency of logistic regression for sex classification. For that it was used dental pieces in a sample of 524 individuals.

Data used

The data used correspond to 524 lower plaster models (264 males and 260 females) of patients treated at an orthodontic clinic in Montevideo, Uruguay (1). From each cast the following were obtained: the mesiodistal diameter, the gingivo-incisal height of both canines and the inter canine distance.

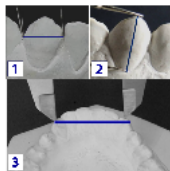


Figure 1: Considered measurements. 1: Mesiodistal Diameter (DMD) 2: Gingivo-incisal Height (AGI) 3: Inter Canine Distance (IC).

Within the set of measurements (in millimeters), multivariate outliers were sought using Mahalanobis distance; observations whose a right or left DMD exceeded 8.5 mm and whose right or left AGI exceeded 11.5 mm were removed, leaving a total of 511 observations.

Measurement	Male (n=255)	Female (n=256)
DMD Right (DMDR)	7.110 (0.440)	6.707 (0.448)
AGI Right (AGIR)	9.231 (0.957)	8.932 (0.951)
DMD Left (DMDL)	7.125 (0.465)	6.744 (0.449)
AGI left (AGIL)	9.260 (0.970)	8.977 (0.959)
IC	26.48 (2.319)	25.491 (2.010)

Cuadro 1: Mean (SD) values by sex for each measurement (without outliers).

Methodology

First a logistic regression model is fitted to discriminate sex by considering all measurements from both the left and the right sides. After that it was used models that only consider the measurements of each side plus the inter canine distance.

- Full Model:
 $Y = \text{Sex}; X = (DMD, DM, DI, AGI, I, DI, C)$
- Left side model:
 $Y = \text{Sex}; X = (DMD, AGI, DI, C)$
- Right side model:
 $Y = \text{Sex}; X = (DMD, AGI, DI, C)$

The study focuses on the predictive power of these models and therefore evaluates their performance through several metrics: Accuracy, Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value, F1 and balanced accuracy... All metrics are calculated by 10 fold cross validation. Analyses are performed in R, mainly using the `caret` library for model fitting and metric extraction.

Results

The estimated coefficients and performance metrics for each model are presented.

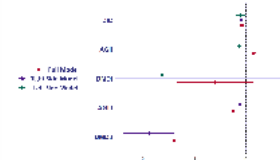


Figure 2: Estimated coefficients and 95% confidence intervals for each model.

Metric	Full	Left side	Right side
Sensitivity	0.678	0.667	0.667
Specificity	0.684	0.668	0.711
PPV (+)	0.681	0.667	0.697
NPV (-)	0.680	0.668	0.682
Accuracy	0.681	0.667	0.697
F1	0.680	0.667	0.681
Balanced ACC	0.681	0.667	0.689

Cuadro 2: One-sided performance metrics.

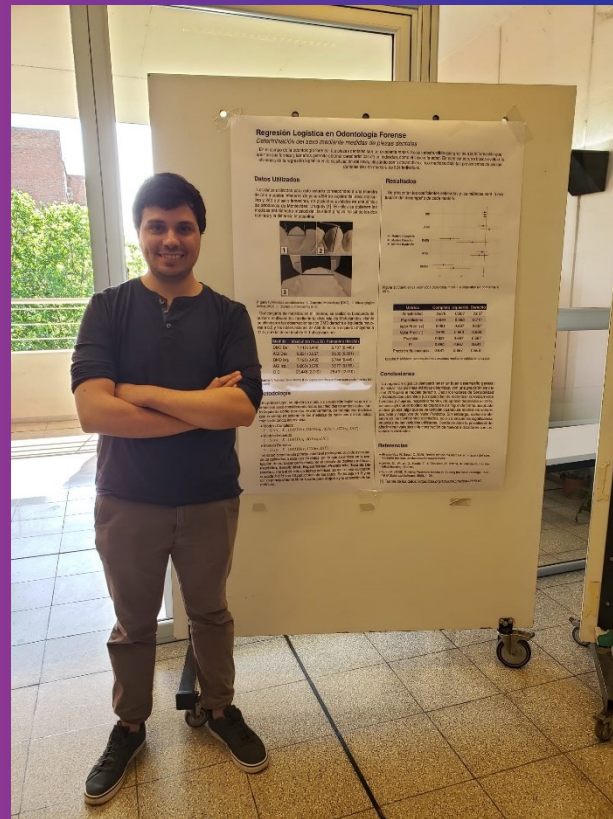
Conclusions

Logistic regression showed good overall performance, with an accuracy of almost 70% for the right side model. Given the sensitivity and Specificity values obtained, both around 70% (the capacity to correctly recognize males and females, respectively) the model can distinguish the two group satisfactorily. The same conclusion is supported by the positive and negative predictive values. However, as seen in the estimated coefficients, only a few of the variables proved statistically significant, suggesting that including additional variables or constructing new indicators from the existing ones could improve the model.

References

- Alvarez-Nar, R. Saez, C. (2020). "¿Indice canino menor? Determinación del sexo mediante técnicas de clasificación supervisada?"
- Arnes, G., Wilken, D., Haeck, T., & Estrella, R. (2013). An introduction to statistical learning (Chilwell).
- Kuhn, M. (2009). Building Predictive Models in R Using the `caret` Package. *Journal of Statistical Software*, 28(1), 1-26.

[1] Data Source: <https://doi.org/10.6089/revista.UUUB9>



UNIVERSITY THIRD PRIZE WINNERS: BRAZIL

Students:
André Augusto
Bechara
Cannizza

Use of Generative AI by Brazilian Economics students: satisfaction and confidence, in the context of learning statistics

Introduction
The use of Generative Artificial Intelligence (AI) has been growing rapidly in several sectors. In education, it allows for real-time explanations of content covered in the classroom, as well as assistance in solving school problems. With this, AI can transform the learning of Statistics, a subject often considered challenging by most students. In this context, in Brazil, the Economics course is one of the ones with the largest number of subjects related to statistics. This reality raises some questions: 1) What is the profile of use of AI by Economics undergraduates? 2) What are their perceptions of General Satisfaction and Academic Confidence with AI in the circumstance of their statistical learning?

Objective
This study aims to analyze the profile of use of generative technologies by Brazilian undergraduate Economics students and their perceptions of General Satisfaction and Academic Confidence with the use of AI, within the scope of statistical learning.

Data
Questionnaire: 10 questions prepared via Microsoft Forms, during the Statistics course, offered in the first semester of 2024. The application took place from 07/06/2024 to 22/06/2024, through a Link made available in the class's virtual classroom. Perceptions of Satisfaction and Confidence were measured on a 5-point Likert scale.

Subjects: 45 students, aged between 19 and 35 years old, in the 4th period of the Economics course at a university located in the State of Pernambuco, Brazil.

Method
Descriptive Statistical Analysis and Non-Parametric Inference were used, employing the Wilcoxon Rank Test and the Kruskal-Wallis Test, with post hoc analyses by way of Bootstrap, ensuring more robust Confidence Intervals for medians.

Frank Wilcoxon (1902-1965) William Kruskal and Milton Wallis (1919-2006) Bradley Jerome (1912-1998) (1938-2004)

All analyses were performed in R (4) Creator of the statistical method.

Results

5.1 Characterization of use

- 87% found out about AI through social media or direct recommendations from friends.
- 91% All students surveyed reported using AI.
- 95% used it to learn Statistics.
- 70% has ever been harmed academically by any erroneous or incomplete information provided by AI.

General Satisfaction Level ★★★★★

Academic Confidence Level ★★★★★

5.2 Insights

A Using the Wilcoxon Rank Test to investigate possible significant differences between students' levels of General Satisfaction and Academic Confidence.

The Test was significant at 2% and the effect size (r) was high, close to 1.00 ($r = 0.908$), suggesting that Satisfaction levels are higher than Confidence levels for most students.

B Application of the Kruskal-Wallis Test to analyze how demographic factors (Gender, Age Range and Income Range) affect students' levels of General Satisfaction and Academic Confidence.

Just the Income presented a significant effect, both on General Satisfaction (p -value = 0.000), and on the level of Academic Confidence (p -value = 0.007). The effect size was high in both cases ($r^2 > 0.340$).

C Post hoc analysis, via Bootstrap⁵⁶, to examine which income range presents a significant median difference in General Satisfaction and Academic Confidence levels.

There was a significant difference in the level of General Satisfaction between students with income range of up to 1 and those with income of 8 to 10 Brazilian minimum wages.⁵⁷

There was a significant difference in the level of Academic Confidence between students with income range of up to 1 and those with income greater than or equal to 11 Brazilian minimum wages.⁵⁸

Final comments
It was found that AI is a consolidated reality for the Economics students surveyed, both in daily use and for statistical learning, confirming the growing acceptance of platforms such as ChatGPT. It was observed that students are satisfied with the technology, however, they demonstrated distrust in its use for statistical learning. This contrast suggests that there is a certain skepticism regarding the effectiveness of AI in certain educational contexts. It was also shown that students with lower family incomes exhibited significantly lower levels of satisfaction and confidence, compared to their higher-income peers. This difference accentuates the inequalities that are so evident in Brazil, and indicates that socioeconomic factors can influence students' perception and experience of new generative technologies.



UNIVERSITY HONOURABLE MENTION: SOUTH AFRICA

Students:
Jaedon Naidu,
Jahnavi Inderjeet

UNLOCKING HAPPINESS: WHAT DRIVES HAPPINESS IN SOUTH AFRICA?

INTRODUCTION

South Africa is a nation blessed for its diversity, with over 48 million people belonging to a variety of ethnic, linguistic, and socio-economic backgrounds. In a world of increasing inequality and digital disruption, how different groups perceive their quality of life. The report that seeks to explore this relationship is quite essential with quality of life, however, it is not enough when the specific needs of the people of South Africa are not considered. This report aims to explore the factors that drive happiness in South Africa, by analyzing the data from the 2014 Survey of Well-being in South Africa. The report will identify the key factors that drive happiness in South Africa, and provide a comprehensive overview of the current state of the country's well-being.

OBJECTIVE

To identify factors that are significantly associated with happiness among South Africans, so that its resources might be allocated effectively to improve the quality of life of its citizens.

THE DATA SET

The General Household Survey by Stats SA is an annual household survey which measures key areas like health, housing, food security, and agriculture, making it ideal for studying socio-economic dynamics in South Africa.

Methodology

- 01. Data Cleaning**
Removed any missing data or non-representative rows in the data.
Converted the variables of interest from a point scale to a binary variable.
- 02. Stepwise Regression**
Used forward selection and backward elimination methods to refine the model.
The approach helped identify the most significant variables to include in the model.
- 03. Random Forests**
Explored Random Forests, a flexible machine learning method that handles complex interactions between predictors without requiring linear relationships.
Assessed feature importance using the Out-Of-Bag (OOB) index.
- 04. Model Refinement**
Performed Gradient Descent with a learning rate of 0.01 for the logistic regression model.
Used Cross-Validation to assess model performance.
Final Output: Selected Random Forest model as the best predictor for happiness.

RESULTS

The most significant variables according to the OOB Gini index:

Variable	Importance
Health	High
Income	Medium
Education	Low

Summary Statistics of the Model:

Metric	Value
AUC	0.85
Accuracy	0.82

CONCLUSIONS

Health as Primary Driver
Healthcare remains a top priority for South Africans, with a strong correlation between health status and happiness. Investments in healthcare infrastructure and services are crucial for improving the overall well-being of the population.

Food Security Matters
Food security is a significant concern, particularly in rural areas. Ensuring access to affordable and nutritious food is essential for improving the quality of life and reducing poverty.

Education Matters
Education is a key driver of happiness, with higher levels of education leading to better employment opportunities and higher income levels. Investing in education is a long-term strategy for economic growth and social development.



ISI World Statistics Congress 2025, The Hague, The Netherlands. This report is a student submission for the University Honourable Mention competition. All rights reserved. © 2025 Jaedon Naidu and Jahnavi Inderjeet.

Students:
Kim Hyunjin,
Park Seungjeong,
Lee Donggeun,
Oh Yujin

Geographical Analysis of regression model with application to Sinkhole occurrence

1. Introduction
A sinkhole is a hole formed when the upper layer of soil collapses due to a void created beneath the ground. In South Korea, frequent sinkholes have occurred, especially in urban areas, and recently, several cases of sinkhole-related fatalities have been reported. Meanwhile, sinkholes often occur near the intersection of roads for pedestrians, and various studies have been conducted to prevent such phenomena. However, since sinkholes occur in South Korea, an annual average of 100 cases, it is necessary to study the occurrence of sinkholes in detail. This study aims to analyze the occurrence of sinkholes in Seoul using Thiessen's method in areas vulnerable to sinkhole formation, considering factors such as urban layout, traffic density, and changes in ground-water conditions. However, installing a lot of devices in the region is expensive.

2. Objectives
The goal is to predict the probability of sinkhole occurrence in order to prepare for and mitigate future sinkholes, and property loss caused by sinkholes. However, the primary causes of sinkhole formation are difficult to investigate, making it difficult to predict the sinkhole occurrence. Therefore, this study aims to predict the probability of sinkhole occurrence for each district in Seoul, the most densely populated city in South Korea. To create a sinkhole occurrence prediction model that incorporates the spatial and environmental characteristics of Seoul, which have not been addressed in previous studies. Through this, the probability of occurrence of sinkholes can be predicted in high-risk districts.

3. Data
The data used in this study is sinkhole occurrence data (Seoul Sinkhole Occurrence Data) provided by the Undergraduate Safety Information Center, and a total of 181 occurrence records.
Sinkhole occurrence data from January 2020 to September 2024.
Building density, traffic volume, building density, population density, residential density, and commercial density were selected as variables expected to affect sinkhole occurrence. In addition, the type of groundwater surface, such as basement surface, building groundwater surface, and construction pit groundwater surface.

4. Analysis
A regression analysis was conducted based on the administrative boundaries of Seoul's districts (Seoul, Paju, Donggi). The model was defined such that Y is the number of sinkholes in a district, and X is the number of independent variables. The corresponding value is Y , otherwise, it is 0. During the analysis, it was identified that the traffic volume variable contained some missing values. Since traffic volume is likely associated with population density, $X_{Traffic}$ variable was calculated using the $X_{Population}$ variable using Thiessen's method. The regression model was then re-estimated using the average traffic volume of the adjacent districts. The objective function was multiplied by the independent variables to create spatial interaction variables, which were incorporated as predictors in the regression model. To compare sinkhole occurrence, logistic regression, and the Zero-Inflated Poisson (ZIP) model, were applied for sinkhole occurrence. The ZIP model and logistic regression were compared using Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) to evaluate model performance. Since accuracy identifies areas where a sinkhole occurs, it is of primary importance, model was considered the most optimal evaluation metric.

5. Result
The figure represents the result of simulating using a sinkhole prediction model. The simulation is applied to the results of the ZIP model. The figure shows the results of the simulation for each district during the simulation. The figure on the right visualizes the simulation results with blue dots representing sinkhole occurrence in each district.

6. Conclusion
The high-risk areas for sinkhole occurrence in Seoul identified by the ZIP model are Gangdong-gu, Yeongdeungpo-gu, Yongsin-gu, Gyeonggi-do, and Gyeonggi-do. The results of the simulation show that the areas of subway stations, high building density, and higher frequency of groundwater usage.
In order to these results, it can be prepared by local governments to control measures of these high-risk areas in the future to prevent sinkhole occurrence.

7. Suggestions
By the primary causes of sinkholes, water leakage from underground water pipes was identified, but data could not be used due to restrictions from local governments. It is necessary to study the causes of sinkholes in detail, and the research could be further advanced.
Since the cause of Seoul's sinkholes is very complicated, it may be necessary to study the differences in sinkhole occurrence based on district, urban layout, spatial characteristics, and topography. In addition, the need to expand the analysis scope to future research using a clustering-based approach.

8. Future work
Sinkholes typically occur in residential areas, but despite an estimated 200,000 to 300,000 cases, height and density of each district, the highest size of sinkhole could be analyzed as follows:
Sinkhole Volume = $\frac{1}{2} \left(\frac{A_{max}}{A_{min}} + \frac{A_{min}}{A_{max}} \right) \times Depth \times \pi$
A model predicting the scale of sinkholes according to the area used in each local government where the largest predicted sinkhole size are expected, which should be taken into account in the future.



THE BEST COOPERATIVE PROJECT AWARD

THE BEST COOPERATIVE AWARD FIRST PLACE

StatBel Academy/Statbel Junior



“With Statbel Academy, Statbel is reaching out to education, and wants to support teachers in their approach. At the same time, future projects are also possible for other target groups. By continuing to focus on statistical literacy, including among children and young people, Statbel wants to contribute positively to a better-informed society in which citizens build up the necessary knowledge to deal critically with data and figures, and so make well-informed decisions.”

THE BEST COOPERATIVE AWARD HONORABLE MENTION

Stats+Stories

STATS + STORIES

A PODCAST ABOUT THE STATISTICS BEHIND
THE STORIES AND THE STORIES BEHIND THE
STATISTICS