MSE 238 Final Project Report

Karththigan Pushparaj (1007143949)

Introduction

Investigating the mechanical properties of alloys and studying the relationships between them is important because many real-world contexts require alloys with properties that are optimized for specific applications. For this reason, this project conducted various statistical analyses, using R, of a Multi-Principal Element Alloy (MPEA) dataset. This dataset includes the values of various mechanical properties for 630 multi-principal element alloys. Another dataset this project analyzed is a Water Quality (WQ) dataset, which presents the levels of *E.coli* detected in the water of various Toronto beaches over a five-month period.

This report presents the results of the analyses used to answer the objectives set in the Project Proposal. These objectives are restated in Table 1.

Table 1: List of Objectives and Corresponding Dataset

Data	#	Objective
MPEA	1	Which categories of materials will have a strength of 500 MPa at 1000K? What is the probability an alloy chosen from the MPEA dataset has these specifications?
	2	Which alloying element in the MoNbTi system has the greatest influence on the overall microstructure of the alloy?
	3	What is the relationship, if any, between the grain size and the yield strength of an alloy? What is the mean grain size distribution in the MPEA dataset?
	4	What is the relationship, if any, between the yield strength and the temperature of an alloy?
WQ	5	Which Toronto beaches have the lowest E.coli levels for the longer period of time?

Objectives 1 and 3 have been slightly modified to allow for greater incorporation of statistical/probabilistic methods. These are discussed in the “Statistical Theory and R modules used” section of this report.

Furthermore, in order to conduct the statistical tests needed to answer these objectives, both the MPEA and WQ datasets were first tidied. This was done using a variety of R packages, which are discussed in the “Statistical Theory and R modules used” section of this report.

Statistical Theory and R modules used

Since several objectives aimed to determine the relationship between multiple properties, the main statistical theory that was used in this project is related to defining the relationship between two or more variables. A qualitative way this was done was using a scatterplot, which plots bivariate data points along two axes. The horizontal axis typically represents the independent variable, and the vertical axis represents the dependent variable. For Objective 3, a scatterplot of yield strength as a function of grain size was produced. A similar plot was produced for Objective 4, but with yield strength as function of temperature.

To quantitatively describe the relationship between yield strength and grain size/temperature, simple linear regression was conducted. This involved determining the equation of the line that best fits the data, which has the form: $\hat{y} = \widehat{\beta_1}x + \widehat{\beta_2}$ [1]. The value of $\hat{y}$ is the predicted yield strength value, x is the grain size/temperature value, and $\widehat{\beta_1}$ is the slope of such a line. Since two variables rarely have a strictly linear relationship, there will be an error associated with the value predicted by this equation. This error is known as the residual, e. It is measured as the distance between the least-squares line and the true y-value of each data point. The line that best fits such data is, then, the values of $\widehat{\beta_1}$ and β_0 that minimizes the value of the squares of the residuals for all the data points. These values were automatically calculated by R's `lm()` function. This project also conducted multiple regression to see if yield strength is a function of both grain size and temperature. Multiple regression is similar to simple linear regression, however $\hat{y}$ is developed as a function of multiple independent variables and has form $\hat{y} = \widehat{\beta_1}x_1 + \widehat{\beta_2}x_2 + \beta_0$. To evaluate the strength of a linear relationship between two variables, the correlation coefficient, r, was calculated. An r value close to 1 indicates a strong linear relationship while a value near zero would suggest no relationship.

For Objective 5, control charts of E.coli levels in beach water were produced. They show the variation in a process over time and the data that is plotted is calculated from a set of subgroups taken over a certain time period. Such data can be the value of the mean, $\bar{x}$, range, R, or standard deviation, S of each sample. This project used $\bar{x}$ −charts to measure variation because the mean level of E. coli is measured. To determine if a process is in control, all the data points in a control chart must fall within the upper (UCL) and lower (LCL) control limits. In an $\bar{x}$- chart, the UCL = $X + A_2\bar{R}$ and LCL = $\bar{\bar{X}} - A_2\bar{R}$, where A_2 is a constant that depends on the subgroup size and $\bar{\bar{X}}$ is the average mean of all subgroups [1].

For Objective 3, histograms and boxplots were used to present the distribution of alloy grain size in the MPEA dataset. Histograms present the frequency distribution of a variable using vertical bars that represent the number of times a variable is within particular range of values. Boxplots are similar to histograms in terms of utility; however, they can be used to identify outliers more easily. The lower edge of a boxplot represents the first quartile, while the top edge represents the third quartile. The lines

extending from the box represent the largest and smallest data points that are within 1.5 times the inner quartile range. Any data point beyond these lines are outliers.

Fundamental probability laws were also used in this project, in particular for Objective 1. The probability of an alloy meeting the criteria in Objective 1 was considered as an event where all outcomes have equal likelihood of occurring. For this reason, the probability of the event that an alloy meets Objective 1 (Event A) can be calculated as $P(A) = k/N$, where k is the number of alloys in Event A and N is the total number of alloys in the dataset. Probabilities for other conditions were calculated in a similar manner but using appropriate probability formulas. For example, the probability that an alloy is multiphase (A) given that it is an MoNbTi alloy (B) was calculated using the conditional probability formula: $P(A|B) = \frac{P(A \cap B)}{P(B)}$.

Table 2 lists the R modules that were used to conduct all these analyses.

Table 2: Main R modules used in this project

Module	Usage
tidyr	This module was used to tidy MPEA dataset into a form where each cell only contains one value. Specifically, the `separate()` function it was used to separate the strings listed in "Phases" column into 3 separate columns titled "Primary Phase", "Secondary Phase", etc.
dplyr	This module was used to was to combine various datasets together. For example, since Objective 2 required investigating the relationship between an alloy's composition, which is listed in the *grouped_by_composition.csv* file, and its phases, which is listed in the MPEA dataset, combining these two datasets was required. This was done using the `left_join()` function.
ggplot2	This module was used to visualize data. Firstly, the `ggplot()` function was used to create a data visualization object. Then, auxiliary "geom" functions were applied to produce diverse types of plots. For example, `geom_point()` was used to create scatterplots and `geom_smooth()` was used to display least-squares lines. Additionally, `geom_boxplot()` was employed to create a boxplot of grain sizes for Objective 3.
stats	The `lm()` function was used to conduct simple and multiple linear regressions. The `quantile()` function was employed to calculate the quartiles of grain sizes of all the alloys in the MPEA dataset. Other similar functions, such as `mean()`, `median()`, etc. were also utilized to produce a statistical; summary of the datasets when required.
qcc	This module's functions were used for Objective 5, where it was necessary to produce control charts. The `qcc()` function was used to create a control chart object and the parameter `type=` was used to specify the types of control chart to be produced (eg. `type=` "S" produces an S-chart.

Results and Discussion

The results of all objectives are presented below, and significant findings are discussed.

Objective 1 :There are 58 alloys that meet these criteria, and these are stored in the variable *objective1* in the R script. To maintain conciseness, only 2/58 alloys are listed here: “Al1 Mo0.5 Nb1 Ta0.5 Ti1 Zr1" ; "Cr1 Mo0.5 Nb1 Ta0.5 Ti1 Zr1". Additionally, the probability an alloy from the MPEA dataset has these specifications was calculated to be $P(meets\ objective\ 1) = 0.09206$.

Objective 2 *:* The probability that an MoNbTi alloy containing a particular additive element is multiphase is listed in the *multiphases_probs* data frame in the R script. It is shown in Table 3 below:

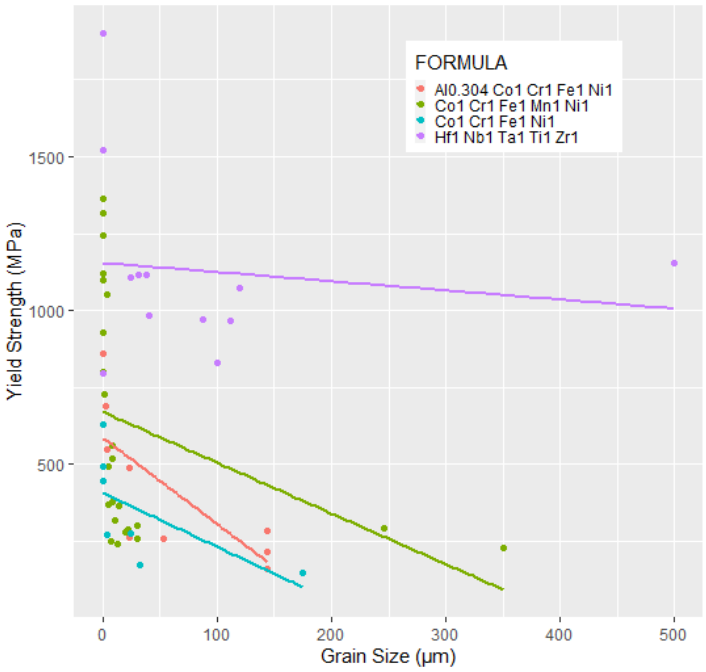
Table 3: Probabilities of various additive elements turning a MoNbTi alloy into multiple phases

Additive Element	**Probability of MoNbTi alloy being in 2+ phases**
Al	0.22
C	0.75
Co	0.75
Cr	0.7
Fe	1
Hf	0.5789474
Si	1
Ta	0.3157895
V	0.1470588
W	0.2857143
Zr	0.2926829

It appears that adding Si or Fe has the greatest impact on microstructure because the probability of an MoNbTi alloy containing these elements being multiphase is 1. However, it is important to note that MPEA dataset does not contain many datapoints for all elements and some additive elements only appear in combinations with other elements in the dataset. This can result in probabilities being much higher than their true value.

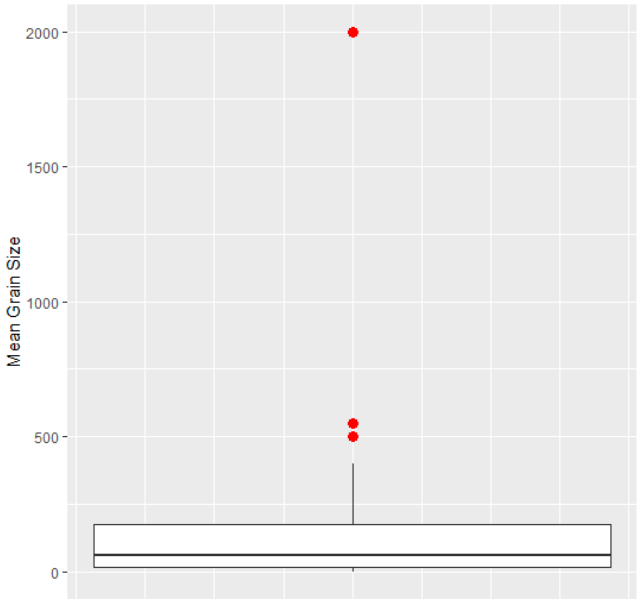
The probability that an alloy randomly picked from the MPEA dataset is multiphase, given that it is an MoNbTi system alloy was calculated as: $P(Is\ Multiphase | Is\ a\ MoNbTi\ alloy) = 0.8636364$

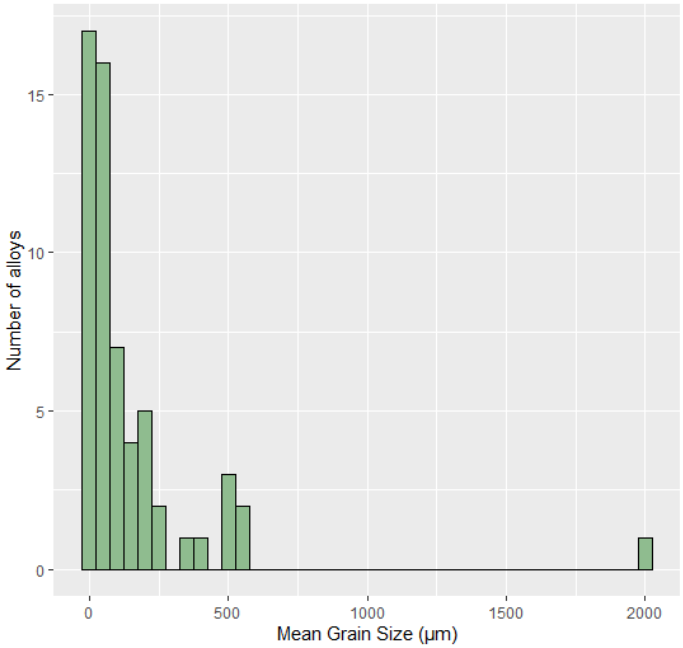
This suggest that most alloys in the MPEA dataset that are from the MoNbTi system are multiphase to begin with.



Objective 3: A scatterplot of yield strength vs. grain size is shown on the right. There appears to be no relationship, according to this MPEA dataset, between grain size and yield strength. This may be because the MPEA dataset does not list grain size values of a single alloy annealed at different temperatures, but rather lists values for multiple alloys of the same composition (i.e., the values are not taken from one alloy undergoing continuous annealing and strength testing but from multiple sources of data).

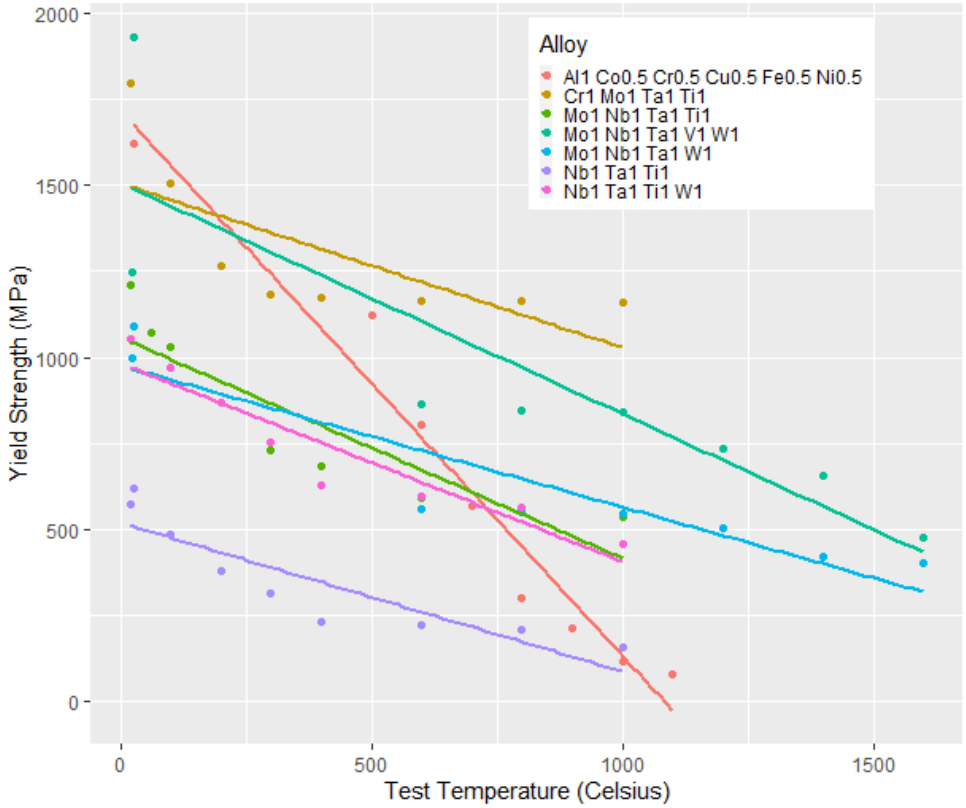
The mean grain size distribution is shown in the histogram and boxplot below. Most alloys have a mean grain size less than five hundred microns. From the boxplot, the median average grain size is 59.325 microns. The red points in the boxplot indicate that three alloys have mean grain sizes greater than 1.5 times the inner quartile range, which indicates that these are outliers.





A multiple regression of yield strength as a function of both temperature and grain size was also conducted. The equation produced was $y = -0.6261x_1 + 0.3450x_2 + 778.4229$, where x_1 *and* x_2 are the temperature and grain size values, respectively. This regression line has an adjusted r^2 value of 0.001581 or an *r*-value of 0.0397. This suggests that there is almost no relationship between yield strength and both temperature and grain size. This unexpected result may be because of the MPEA dataset's limitations, which were mentioned above.

Objective 4: A scatter plot of yield strength vs temperature for various alloys is shown below:



Clearly, yield strength decreases as temperature increases. This is proven by the negative slopes of all regression lines. The slopes of these lines, in addition to the values of the correlation coefficient for each least-squares line of each alloy, were calculated (in the *lm_coeffs* data frame) and are listed in Table 4.

Table 4: Yield strength vs. temperature linear regression details of each alloy

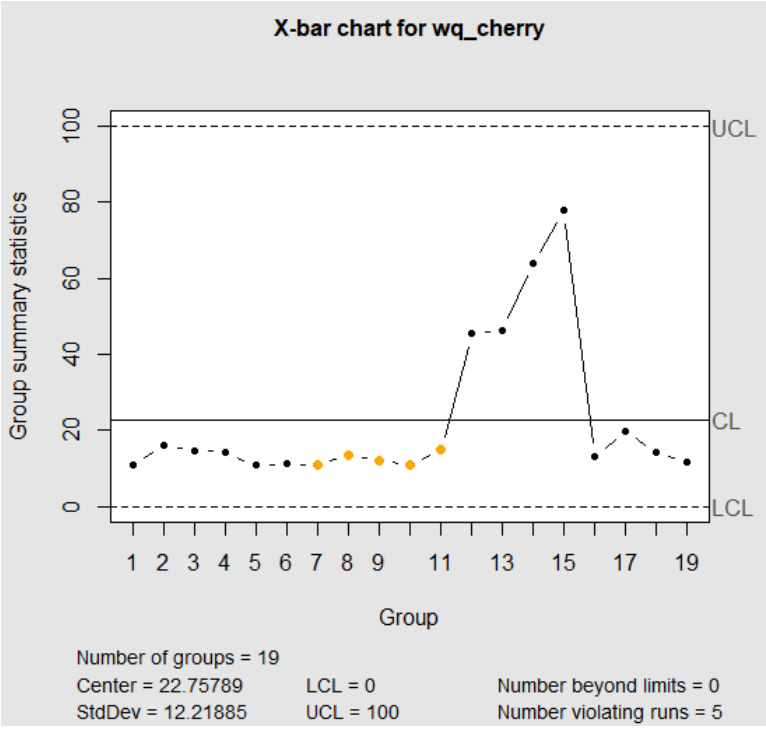
Formula	**Slope**	**Y-intercept**	**r^2**	**r**
Al1 Co0.5 Cr0.5 Cu0.5 Fe0.5 Ni0.5	-1.581	1714.605	0.960	0.980
Cr1 Mo1 Ta1 Ti1	-0.476	1504.472	0.505	0.711
Mo1 Nb1 Ta1 Ti1	-0.641	1058.082	0.817	0.904
Mo1 Nb1 Ta1 V1 W1	-0.670	1506.145	0.764	0.874
Mo1 Nb1 Ta1 W1	-0.410	975.943	0.868	0.932
Nb1 Ta1 Ti1	-0.431	520.063	0.798	0.893
Nb1 Ta1 Ti1 W1	-0.575	982.122	0.891	0.944

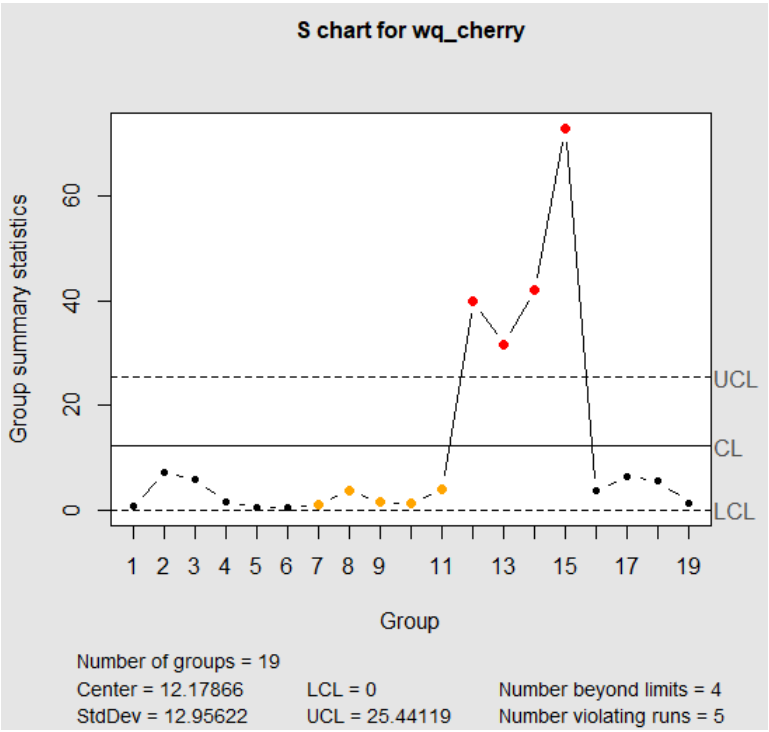
The *r* values of all the alloys' regression lines are close to 1, which indicates that the relationship between yield strength and temperature is roughly linear.

Objective 5: X-bar charts with a subgroup size of 5 days and an upper control limit of 100 E.coli were used. This is because in Toronto, E.coli levels greater than 100 are deemed unsafe for swimming [2].

Beach	Number of Points Above UCL in S chart	Number of Points Above UCL = 100 in X-bar chart
Bluffer's Beach Park	3	3
Centre Island Beach	2	1
Cherry Beach	4	0
Gibraltar Point Beach	1	0
Hanlan's Point Beach	1	0
Kew Balmy Beach	3	3
Marie Curtis Park East Beach	2	5
Sunnyside Beach	3	6
Ward's Island Beach	2	1
Woodbine Beaches	2	2

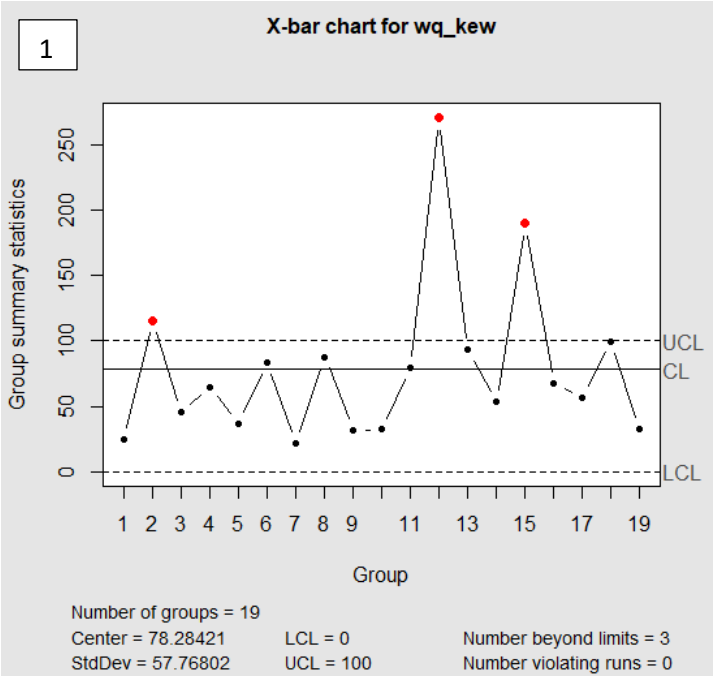
Clearly, the safest beaches are Cherry, Gibraltar, and Hanlan's Point beach because they do not have any periods where the 5-day average E.coli level is greater than 100, suggesting that their E.coli variation is in control. The x-bar chart of Cherry beach is shown below. However, its S-chart shows that its variability in E.coli levels is not predictable (i.e., not in control).

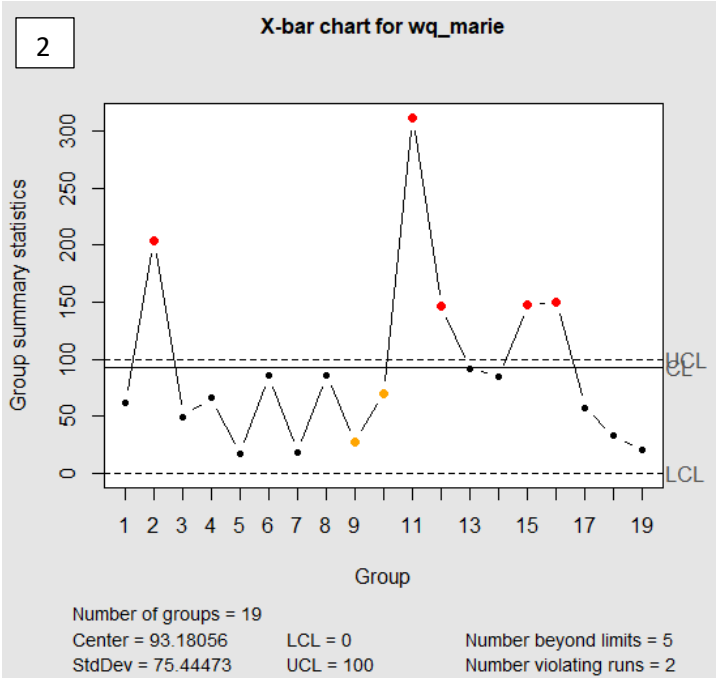


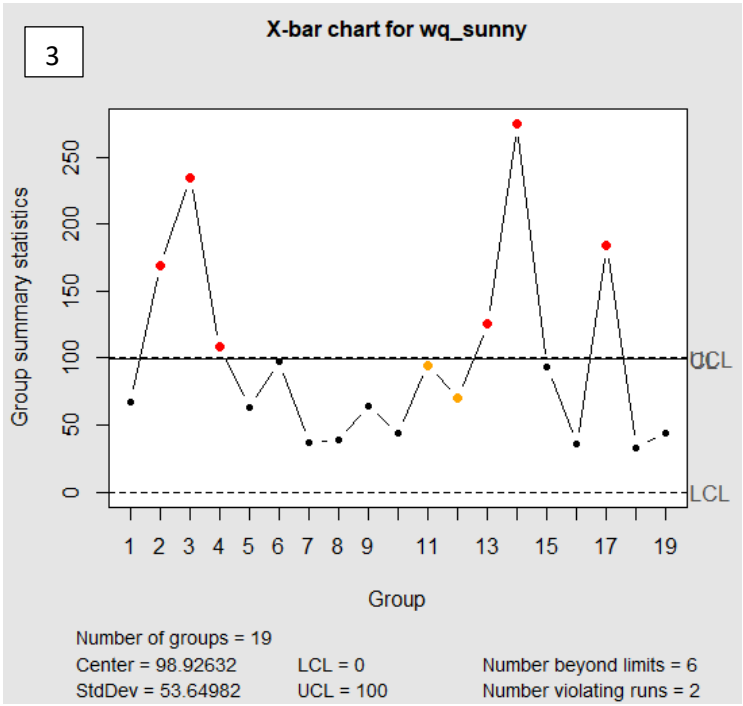


Overall, however, Gibraltar and Hanlan's Point Beach are the safest beaches for swimming because 1) they have the fewest points above the UCL in the S-chart, suggesting that their E.coli fluctuation is predictable and 2) they have zero 5-day periods where the mean E.coli level is above the UCL in the x-bar chart, suggesting that their levels are normally less than what is considered dangerous for swimming.

In contrast, Marie Curtis Beach, Kew Balmy Beach, and Sunnyside Beach are the most dangerous beaches to swim in because the central control limits in their x-bar charts are very close to the UCL of 100, suggesting that the E.coli levels at these beaches are typically high enough to be harmful. These charts are shown below; 1) is Kew Balmy Beach, 2) is Marie Curtis beach and 3) is Sunnyside Beach.







Conclusion

The results from this project demonstrate that certain mechanical properties of multi-principal element alloys have relationships that can be meaningfully quantified using standard statistical and probabilistic tools. Such was the case with yield strength and temperature, which share a negative linear relationship as determined by simple linear regressions. Similarly, probabilistic analysis was used to determine the influence of specific additive elements on the overall microstructure of an MoNbTi-system alloy.

On the other hand, the relationship between grain size and yield strength could not be meaningfully extracted from this data set, despite it being known that increasing grain size typically lowers yield strength [3]. This suggests a potential limitation of the MPEA dataset: since it is a compilation of data from published articles [4] and not data produced from an experiment, it may not be easy to study continuous relationships. In other words, many different papers can report different grain size values for the same alloy at a single yield strength and this would appear as multiple observations in the dataset, whereas, in reality, an alloy will only have one grain size value. This would result in poor linear fitting when a linear regression is conducted. In contrast, the Water Quality dataset was sufficiently compiled for this project to validly determine which beach in Toronto is the safest to swim in.

There are several possible extensions for this project. One extension may include finding different regressions fits between properties other than a simple or multiple linear regression fit (e.g., polynomial, rational, etc.). These different fittings can then be compared to each other to determine which type of regression most accurately reflects true relationships between properties. Another extension is to include probabilistic analysis of the Water Quality dataset.

In conclusion, the findings of this project demonstrate that statistics and probability are powerful tools that can be used to further expand materials science and engineering.

References

[1] W. Navidi, ISE Principles of Statistics for Engineers and Scientists, 2nd ed. New York: McGraw-Hill Education, 2021.

[2] C. Toronto, C. People, H. Wellness, H. Monitoring, B. Quality and A. Quality, "About Beach Water Quality", City of Toronto, 2022. [Online]. Available: https://www.toronto.ca/community-people/health-wellness-care/health-inspections-monitoring/swimsafe/beach-water-quality/about-beach-water-quality/. [Accessed: 03- Apr- 2022].

[3] S. Whang, "Introduction", *Nanostructured Metals and Alloys*, p. xxi-xxxv, 2011. Available: 10.1016/b978-1-84569-670-2.50028-9 [Accessed 4 April 2022].

[4] C. Borg et al., "Expanded dataset of mechanical properties and observed phases of multi-principal element alloys", *Scientific Data*, vol. 7, no. 1, 2020. Available: https://doi.org/10.1038/s41597-020-00768-9. [Accessed 4 April 2022].