Final Competition Information & Guidelines

Data Science and Machine Learning Club

FEBRUARY 2023

Introduction & Background

The Data Science and Machine Learning Club at the University of Calgary is hosting a competition to give its students an opportunity to showcase their skills in programming and data science. This competition is open to students enrolled in the university, and it provides a platform for them to demonstrate their abilities to the wider community. The competition organizers have provided a comprehensive dataset that the participants will work with.

The objective of the competition is for the students to perform a variety of data science tasks, including visualization, manipulation, and analysis. This will require the students to use their programming and data analysis skills to uncover insights and trends within the data. The participants will have the opportunity to apply the concepts and techniques they have learned in the classroom to real-world data, and to demonstrate their ability to work with large datasets.

In addition to demonstrating their skills, the competition provides an opportunity for students to gain recognition for their accomplishments. This can be valuable for students who are looking to build their portfolios and further their careers in programming and data science.

Overall, the competition aims to foster a sense of community among University of Calgary students and to provide an avenue for them to demonstrate their skills to the world. By participating in this competition, students can gain valuable experience and make connections that can help them achieve their career goals.

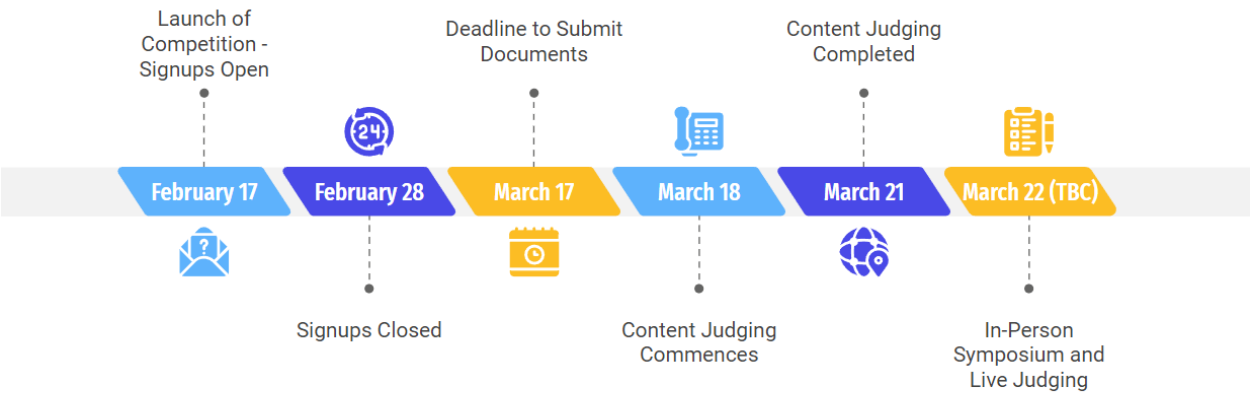
Competition Categories

The competition is structured into three distinct categories, allowing participants to compete in the area that showcases their skills and interests. These categories are:

- **Data Visualization:** This category focuses on the visual representation of data and aims to test the participants' ability to create compelling and informative visualizations that communicate insights and patterns in the data. Participants will be expected to use their knowledge of data visualization techniques and tools to produce high-quality visualizations.
- **Machine Learning Approach:** This category focuses on the application of machine learning algorithms to the data. Participants will be expected to demonstrate their ability to clean and pre-process data, select appropriate algorithms, and evaluate their results. They will also be expected to use their skills in programming and data analysis to implement and apply machine learning techniques to the dataset provided.
- **Research Proposal Approach:** This category focuses on the formulation and presentation of a research proposal based on the data provided. Participants will be expected to use their skills in data analysis and critical thinking to identify important questions and trends in the data, and to propose a research question(s). They will also be expected to present their findings in a clear and concise manner, demonstrating their ability to communicate complex ideas effectively.

Each category offers its own unique challenges and opportunities for participants to showcase their skills and knowledge. Participants are **only allowed to compete in one category,** and the **category must be declared at the time of submission**. The competition provides a valuable learning opportunity for students who are interested in advancing their careers in data science and related fields.

Timeline



Relevant Participant Information

This statement refers to the rules and regulations of a competition.

- The competition allows participants to collaborate and work with a partner, but they must declare this when they sign up for the competition. If they fail to do so, they may be disqualified.
- You cannot recreate the figures present in the WHR report. All figures generated must be unique, and any references to the WHR report must be cited.
- The competition also allows participants to use additional open source datasets to supplement the data provided by the organizers. However, these datasets must be open access and must be relevant to the category they

have chosen to participate in. The participants must reference these additional datasets in their submissions.

- Additionally, it should be noted that participants are only allowed to use a maximum of two (2) open source datasets as supplementary data. The use of more than two datasets may result in disqualification. It is important for participants to carefully choose the datasets they use and make sure that they align with the category they have chosen and aid the provided data, as per the rules and guidelines set by the competition organizers
- It is important to note that the purpose of using the supplementary data must be to aid the provided data, and not to be the main focus of the study. The judges of the competition have the discretion to determine if the use of supplementary data was excessive and may result in disqualification. Therefore, participants are encouraged to review the example section of the competition rules and guidelines for guidance on how to properly use additional data.

About the Dataset

The World Happiness dataset is an annual survey conducted by the Sustainable Development Solutions Network that measures the subjective well-being of people living in different countries. It is based on factors such as income, social support, freedom, trust, generosity, and health, among others.

The dataset contains information on 156 countries, including:

- Country: The name of the country
- Year: The year of the survey
- Life Ladder: A measure of overall life satisfaction on a scale from 0 to 10
- Log GDP per capita: The natural logarithm of the country's GDP per capita, adjusted for purchasing power parity
- Social support: A measure of perceived social support on a scale from 0 to 10
- Healthy life expectancy at birth: The number of years a newborn can expect to live in good health, taking into account both mortality and morbidity
- Freedom to make life choices: A measure of perceived freedom to make life choices on a scale from 0 to 10
- Generosity: A measure of perceived generosity based on donations in the past month on a scale from 0 to 10
- Perceptions of corruption: A measure of perceived corruption in government and business on a scale from 0 to 10

These variables can be used to study the factors that contribute to overall happiness levels in different countries, and to develop solutions that can help improve the well-being of individuals and societies worldwide. Due to changing organizations that collect this data, the variables may differ from year to year.

Data Visualization Award Guidelines

Overview of Award

Data visualization plays a crucial role in data analysis and understanding. It helps to effectively communicate complex data patterns, relationships, and insights to a wide audience, making it a valuable tool in the field of data science. Data visualization allows individuals to gain insights and make informed decisions based on the data being presented. In the context of a competition, it is used to showcase the participants' understanding of the provided data and their ability to effectively communicate their findings. The guidelines for data visualization in this competition provide participants with the necessary tools and resources to create meaningful and impactful visualizations.

The competition organizers **recommend participants in this category to use the programming language Python 3.0** and specific libraries, such as Matplotlib, Plotly, and Seaborn, to create visualizations of the provided data. However, **participants are not restricted to using only these tools** and are free to choose any other programming language they are comfortable with.

The organizers acknowledge that programming is not a requirement in the field of data science and therefore, **participants are allowed to use external software tools, such as Excel, Google Sheets, Tableau, or others**, to design their figures and visualizations. However, it is important for participants to ensure that they have proper permissions to use these external tools and to follow the rules and guidelines set by the competition organizers.

Submission Criteria

- Submission deadline will be March 17th at 11:59 PM MST. Any submissions past this deadline will not be considered for the competition.
- Participants are required to submit the following documents:

- Source code as .py or other appropriate format based on the chosen programming language. If participants choose to use external softwares such as Excel or Tableau, then they must submit all relevant files.
    - If an external software is used, please submit a one-page document indicating the work process undertaken to design the visualizations.
- Document containing the visualizations in either of the following formats:
    - A 16:9 poster presentation with all the figures and relevant notes
    - A virtual display presentation, including but not limited to a GUI, a Website link or other similar formats. Participants choosing to present data through a virtual format must submit a one-page instruction guide alongside their code.

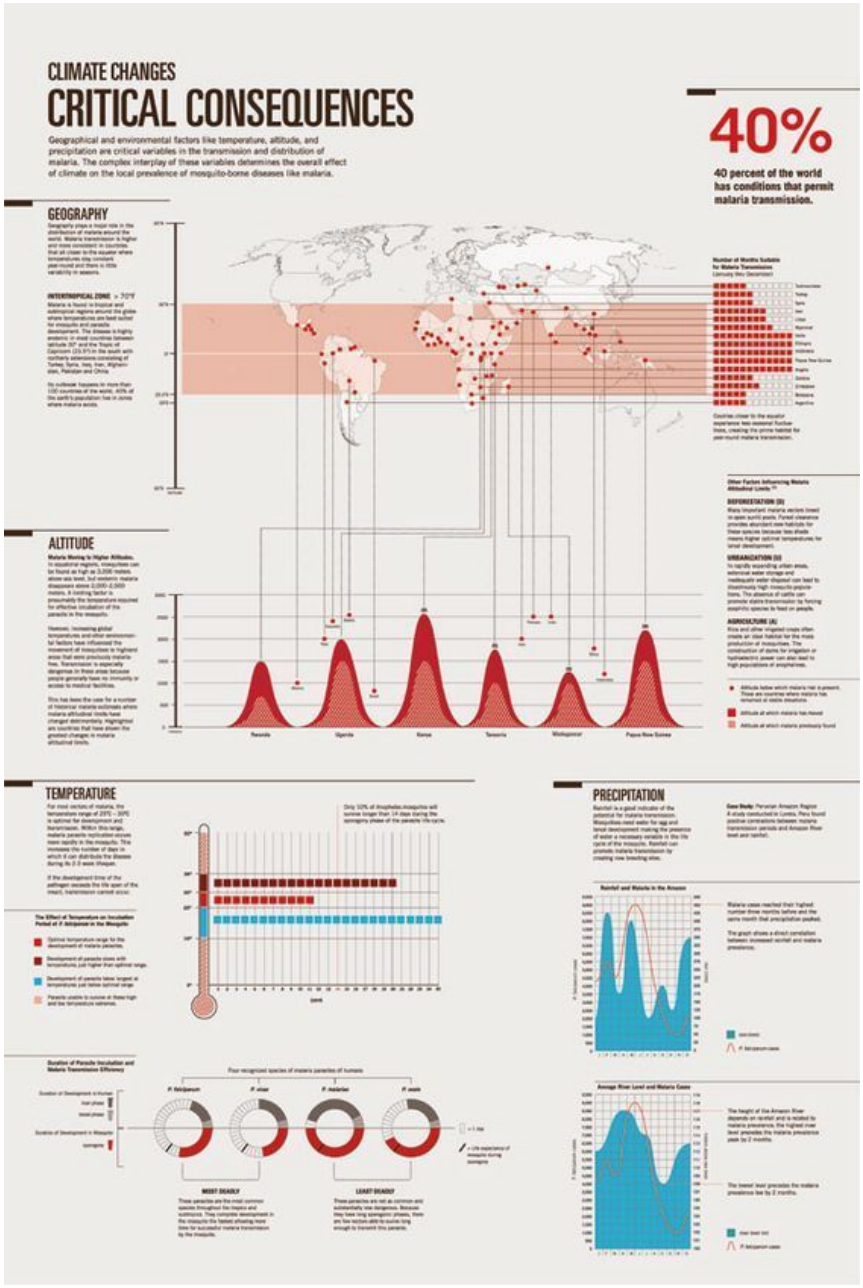
Examples

This is an example from Kaggle, which walks through some of the common ways to create graphs in python. Please use it as a reference on how your code should be annotated and structured (if you decide to incorporate coding into your project).

https://www.kaggle.com/benhamner/python-data-visualizations

Alternatively, if you choose to use softwares such as Google Sheets, Excel or Tableau, please submit the relevant files.

If you choose to present a poster, the next page shows an ideal example of what content would be presented.



Important Notes/Considerations

Please think creatively when generating your visualizations! It is important to have unique but relevant graphs and depictions.

Machine Learning Award Guidelines

Overview of Award

This statement refers to a specific award in the competition, which will be given to the participant(s) who use machine learning in their program. The program must use the provided data to categorize, predict, or make a decision, and participants have the freedom to choose what the machine learning system does. Due to the vast applications of machine learning, participants have full creative freedom within this category as long as machine learning is implemented.

Participants are allowed to use **any type of machine learning application**, such as KNN, GANs, Linear Regression, Logistic Regression, etc., in their program, as long as the code used is original. This allows participants to showcase their knowledge and skills in the field of machine learning and demonstrates their ability to apply these concepts to the provided data. The use of original code also ensures that the competition remains fair and that participants are not using pre-existing solutions or tools.

In summary, the award is given to participants who use machine learning in their program and successfully apply it to the provided data, using original code and demonstrating their knowledge and skills in the field.

Submission Criteria

- Submission deadline will be March 17th at 11:59 PM MST. Any submissions past this deadline will not be considered for the competition.
- Participants are required to submit the following documents:
    - Source code as .py or other appropriate format based on the chosen programming language.
    - Document containing the visualizations in either of the following formats:

- A 16:9 poster presentation with all the figures and relevant notes
- A virtual display presentation, including but not limited to a GUI, a Website link or other similar formats. Participants choosing to present data through a virtual format must submit a one-page instruction guide alongside their code.
- A video demonstration of the designed program. Video files must be in either QuickTime or MP4 format when submitted. The video may not exceed 2 minutes in length. Participants may be asked to provide a live demo of the program on the day of the symposium upon the request of the Judges.

Examples

Listed below is an example of linear regression done using Python and other relevant libraries. This is a good template on how your code can be annotated, and what comments you choose to include in your code file. It also gives a general example as to how you can use the provided data in a machine learning focused approach.

https://www.kaggle.com/goyalshalini93/car-price-prediction-linear-regression-rfe

This image in the next page is an example of how you can present your findings using a 16:9 poster. It contains an overview of your approach and all relevant findings. It is also aesthetically appealing.

Important Notes/Considerations

Please think creatively when generating your visualizations! It is important to have unique but relevant graphs and depictions.

Listen to Your Data: Turning Chemical Dynamics Simulations into Music

Austin Atsango[1] (atsango), Soren Holm[1] (sorenh), and K. Grace Johnson[1] (kgjohn)

[1] Department of Chemistry, Stanford University

Abstract

Our goal is to translate simulation data into a musical form in order to present a different way to interact with data. Specifically, the goals are 1) to **generate music**, i.e. melodies that are indistinguishable from those composed by humans, and 2) to have those melodies **reflect trends in the underlying data**.

We take two approaches: 1) We use a supervised model (either **softmax regression** or an **LSTM RNN** trained on composed melodies) to predict the next note in a song, biased by the trajectory values. 2) We cluster snippets of a trajectory using a Gaussian Mixture Model (**GMM**) with the EM algorithm to discover motifs within a trajectory, then match these motifs to similar ones from a composed melody.

We evaluate the success of these approaches with a survey designed to assess the two goals of the project.

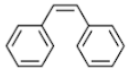
Datasets

Music

- MIDI data format
- 312 classical piano pieces
- 93 piano pieces from Final Fantasy video game
- Simplified using music21 and mido packages in Python to represent as pitch (with value 0 to 127) vs time

Most piano pieces have melodies in pitch range 50-90

Chemical dynamics



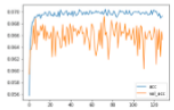
(Z)-Stilbene

- Quantum dynamics simulations of stilbene decaying from excited to ground state
- 200 trajectories of potential energy vs. time (femtoseconds)
- Potential energy normalized to 50-90 pitch range

Models

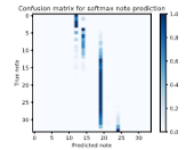
Predictive models

- Convolution over each musical piece
- One-hot encoding
- Supervised: predict next note based on previous 50 notes

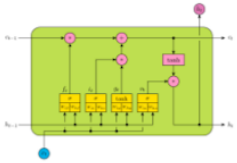


Softmax Regression

$$l(\theta) = \sum_{i=1}^{n} \log \prod_{l=1}^{k} \left(\frac{\exp(\theta_l^T x^{(i)})}{\sum_{j=1}^{k} \exp(\theta_j^T x^{(i)})} \right)^{1\{y^{(i)}=l\}}$$



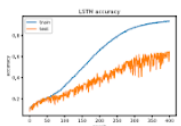
LSTM RNN

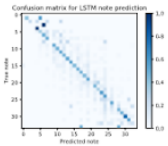


$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$
$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$
$$C_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_C)$$
$$o_t = \sigma(W_o [h_{t-1}, x_t] + b_o)$$
$$h_t = o_t \times \tanh(C_t)$$

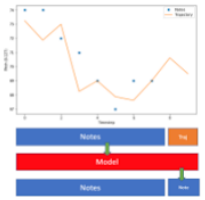
Architecture:

- LSTM with 256 hidden units
- LSTM with 38 hidden units
- Dense layer with softmax activation



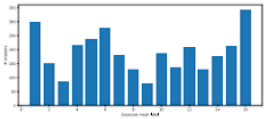


Music generation from trajectory:



GMM

- use a Gaussian Mixture Model with the EM algorithm to cluster snippets of all trajectories based on distance and gradient
- Match snippets to motifs in a given musical piece



Population of trajectory snippets for EM with 16 GMM means

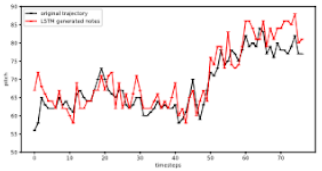
Results

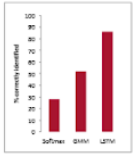
Goal 1: Turing test

Generated	
LSTM full	56%
LSTM subset	64%
Generated based on trajectory	
Softmax regression	13%
GMM	45%
LSTM subset	42%
Real music (control)	57%

Survey results with 40 participants showing % responding the sample was composed by a human.

Goal 2: matching generated music to trajectory





Percentage of participants matching correct trajectory

Discussion

We explored training predictive models with several architectures and on several subsets of the music data. We found the best training and validation accuracy using a subset of the full dataset: the pieces composed by Clementi.

Of all models tested, the LSTM RNN was most successful at generating music that reflected trends in a given dynamics trajectory. Softmax regression produced samples with the same note repeated, which were neither musical nor reflective of trajectory data. The GMM approach had roughly the same success as the LSTM, but cannot truly be considered music generation, as it sampled snippets from composed pieces. To more fully analyze the success of the models in achieving both goals outlined, we would need a survey with a much larger sample size both in number participants and number of audio clips.

Future Work

Future efforts include curating a larger dataset with distinctive melodies and exploring other generative models such as GANs or GRUs. The control of the Turing test shows that reducing a piece to simply pitch and time removes much of the musicality. We would also want to extend the model to train not just on pitch, but also on rhythm, chords, and other expressive information, then explore methods of interpreting the trajectory data with these additional features.

References

[1] Weir, H., Williams, M., Parrish, R., and Martinez, T.J. Nonadiabatic dynamics of photoexcited cis-Stilbene using ab initio multiple spawning. *In prep.* (2019).

[2] Classical piano midi page. Retrieved from http://www.piano-midi.de/

[3] Skli, Sigurur. How to Generate Music Using a LSTM Neural Network in Keras. Data Science, 7 Dec. 2017. Retrieved from towardsdatascience.com

[4] Holzner, A. LSTM Cells in Pytorch. Retrieved from https://medium.com/@andre.holzner/lstm-cells-in-pytorch-fab924a78b1c

Research Approach Award Guidelines

Overview of Award

The award competition described in the statement is a research-based challenge where participants are tasked with using data to answer a question of their own design. The data provided as part of the competition is the starting point for participants to generate a research question that is reflective of the information contained within the data.

The scientific method is a systematic approach to solving problems and making decisions that involves several steps including formulating a hypothesis, designing an experiment to test the hypothesis, collecting and analyzing data, and interpreting the results. Participants are expected to follow these steps in their research to come up with a valid and meaningful conclusion.

The use of statistical analysis or other relevant analytical approaches is critical to the competition. Participants must use these methods to analyze the data and obtain results that will help answer their research question. The goal of this competition is to encourage participants to use data to drive their research and make informed decisions based on their findings.

It is worth noting that the sample questions listed are provided as examples and cannot be used in the competition. Participants must come up with their own research question, reflecting the data provided, that they can then work on answering using the scientific method and statistical analysis. The objective is for participants to show their creativity and problem-solving skills by designing their own research questions and using the data provided to answer them in a meaningful and scientifically sound manner.

Submission Criteria

- Submission deadline will be March 17th at 11:59 PM MST. Any submissions past this deadline will not be considered for the competition.

- Participants are required to submit the following documents:
    - Source code as .py or other appropriate format based on the chosen programming language.
    - Document containing a 16:9 poster presentation with all the figures and relevant notes.
    - Manuscript explaining the research question, methodology, results, conclusions and references. The manuscript will not exceed more than two (2) pages (not including references).

Examples

This is an example from the Kaggle website on how to conduct a statistical analysis using Python. Please refer to it when generating your notebook, and annotate your code accordingly.

https://www.kaggle.com/kanncaa1/statistical-learning-tutorial-for-beginners

Please refer to the next page to observe a poster presentation done by previous students. Please take note of the dimensions, as well as content organization within the poster.

Important Notes/Considerations

Please follow the scientific process when conducting the statistical research approach. Only a poster/infographic style document can be submitted.

Mutations within DPP9, IFNAR2, SLC6A20, and TMPRSS2 Associated with COVID-19, But Show No Distinct Correlations With COVID-19 Severity

Introduction

This study aims to determine the possible correlation between mutation frequency in the human genes DPP9, IFNAR2, SLC6A20, and TMPRSS2 and mortality rate for COVID-19.

Sars-CoV-2, also known as COVID-19, is a virus that attacks the human respiratory systems and can lead to death in serious cases. Scientists have noted that mutations in each of the following genes may be associated with COVID-19; however, the data has not yet been thoroughly analyzed and distinct correlations between mutation frequencies and severe COVID-19 outcomes have yet to be established.

Gene	Function in relation to COVID-19
DPP9	DPP9 is a dipeptidyl peptidase that works to cleave off the N-terminal dipeptides of proteins in order to activate them. DPP9 plays a role in lung inflammation during COVID-19, which suggests a possible link to COVID-19 severity.
INFAR2	IFNAR2 encodes interferons that, when bound to ligand, starts phosphorylation cascades involved in the production of proteins involved in the immune system. This is important for the development of receptors that stimulate the immune system in the event of a virus like COVID-19.
SLC6A20	SLC6A20, which codes for proteins involved in transport substrates, relates to sever COVID-19 outcomes as it the amino acid transporter it codes for interacts directly with ACE2. ACE2 is the primary receptor used by COVID-19 to enter the body.
TMPRSS2	TMPRSS2 is involved in the production of cell-surface proteins that line both the digestive tract and the lungs. COVID-19 may need the protein produced in order to enter the body

Results and Analysis

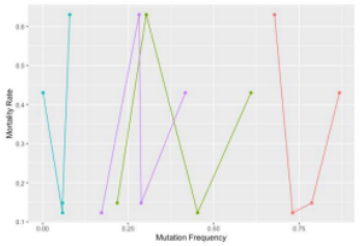


Fig 1: Line plot of Allelic Mutation Frequency vs. Mortality Rate

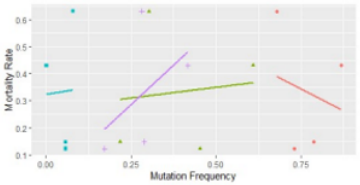


Fig 2: Linear Regression of Allelic Mutation Frequency versus Mortality Rate correlation for DPP9, IFNAR2, SLC6A20, and TMPRSS2.

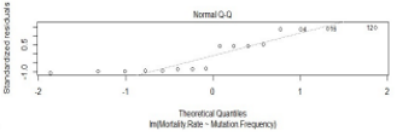


Fig 3. Normal Q-Q plot for Shapiro-Wilk Test performed on correlations to confirm the non-normal distribution of results. The p-value is 0.002714 and the points deviate from the line a fair amount, which indicate non-normal distribution.

Gene	Rho-value
DPP9	-0.2
IFNAR2	0.0
SLC6A20	+0.32
TMPRSS2	+0.4

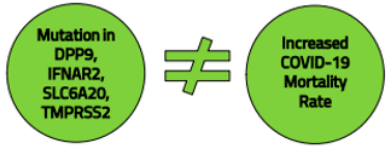
Table 1: rho-value for each mutation frequency in relation to mortality rate after a Spearman Correlation. ***Note: since all of the data was obviously not normally distributed (based on figure 1), a non-parametric test (ie. Spearman Correlation) was used for analysis.*

Analysis:
Based on Fig 2 and Table 1, it is clear that the relationship between allelic mutation frequency and mortality rate is not significant for any of the genes. The rho-values are much lower than the desired |0.7| or greater value for significant correlation.

Discussion

No statistical significance seen in the relationship between mutation frequency (for these 4 genes) and mortality rate

Figure 2 and Table 1 (linear regression via Spearman correlation analysis) show a flat slope for the trendlines and low rho-values (below |0.7|), which indicates the lack of significance



The genes studied don't have a role in COVID-19's interaction with the human body

- Though DPP9 plays a role in lung inflammation, it may not directly interact with the COVID-19 virus
- The proteins coded for by IFNAR2 may play a role in immune function unrelated to COVID-19
- While SLC6A20 encodes transport proteins that interact with the receptors used by viruses to enter human cells, the indirect nature of its interaction with the COVID-19 virus may not result in a strong impact on mortality
- TMPRSS2 may not encode for proteins that facilitate the entry of COVID-19 into host cells (unlike other viruses)

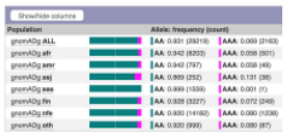
Various limitations (ie. small sample size, generalizations made for comparing mutation frequency to mortality rate) and confounding variables may have impacted results

A more extensive study would be beneficial for further research and increased accuracy in results; future research may also include studying a greater variety of genes

Methods

Step 1: Conduct background research to determine specific genes that could affect the severity or mortality rate of COVID-19.

Step 2: Collect mutation frequency data for selected genes and COVID-19 mortality rate data for respective ethnicities. Data collected from 1000 Genome Project (via Ensembl) and CDC COVID Tracker.



Step 3: Compile all genetic and COVID-19 data into one table.

Step 4: Run Shapiro-Wilk Test to confirm non-normal distribution. Then run Spearman correlation to determine if correlations exist between mutation frequency and mortality rate.

Conclusion

The mutation frequencies of all four genes (DPP9, IFNAR2, SLC6A20, TMPRSS2) are not statistically strongly correlated with COVID-19 mortality rate (as determined by linear regression analysis via Spearman correlation)

The genes in question are not associated with severity of COVID-19

The hypothesis, which stated that there would be a strong correlation between mutation frequency and mortality rate, cannot be accepted

- Predicted that DPP9 and IFNAR2 would display positive correlation → neither displayed a strong correlation, DPP9 showed a negative trend while IFNAR2 showed a positive trend
- Predicted that SLCA60 and TMPRSS2 would display negative correlation → neither displayed a strong correlation, both show positive trend

Acknowledgements

1.Bjelke, J. R., Christensen, J., Nielsen, P. F., Branner, S., Kanstrup, A. B., Wagtmann, N., & Rasmussen, H. B. (2006, June 01). Dipeptidyl peptidases 8 and 9: Specificity and molecular characterization compared with dipeptidyl peptidase IV. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1462722/

2.Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., . . . MacArthur, D. G. (2020, May 27). The mutational constraint spectrum quantified from variation in 141,456 humans. Retrieved from https://www.nature.com/articles/s41586-020-2308-7#citeas

3.Kozlov, M. (2020, December 14). Key Genes Related to Severe COVID-19 Infection Identified. Retrieved from https://www.the-scientist.com/news-opinion/key-genes-related-to-severe-covid-19-infection-identified-68276

In-Person Symposium

The competition will culminate in an in-person symposium, where participants will have the opportunity to present their work to a panel of judges and an audience of their peers, industry professionals, and members of the academic community. The symposium is scheduled to take place on March 15th, and it promises to be an exciting and enlightening event for all who attend.

During the symposium, participants will have the opportunity to showcase the results of their work and to explain their methods and findings. The presentations will be evaluated by a panel of experts in the field, who will assess the quality of the participants' work and provide feedback on their projects. Awards will be given to the top performers in each category, recognizing their exceptional skills and achievements.

The symposium is also an opportunity for students to network with industry professionals, who may provide valuable insights into the latest developments in the field and help participants establish connections that can further their careers. The in-person symposium provides a platform for participants to demonstrate their abilities, to receive recognition for their achievements, and to gain valuable experience in presenting their work to a wider audience. It is an exciting and important event for all those who have participated in the competition and will be a testament to their hard work and dedication.

Judging and Evaluation Criteria

Data Visualization Award Criteria

These criteria provide a comprehensive evaluation of the data visualizations and the ability of the participant to demonstrate their proficiency and creativity in the field of data visualization. The final decision on the award winner will be based on a thorough assessment of all the criteria.

- **Creativity:** The originality and creativity of the visualizations and their ability to effectively convey insights and patterns in the data.

- **Clarity:** The clarity of the visualizations and their ability to effectively communicate information and findings to a non-technical audience.

- **Relevance:** The relevance of the visualizations to the provided data and the problem being addressed.

- **Technical Skill:** The demonstration of technical skill in the creation of the visualizations, including the use of programming languages and libraries such as Python, Matplotlib, Plotly, and Seaborn, or external software such as Excel, Google Sheets, and Tableau.

- **Data Analysis:** The ability of the participant to perform a thorough analysis of the provided data and to effectively use the results to inform their visualizations.

- **Color Usage:** The effective and appropriate use of color in the visualizations to convey information and enhance the overall aesthetic appeal of the visualizations.

- **Layout:** The design and layout of the visualizations, including the arrangement of charts and graphics, the use of space, and the overall presentation of the information.

- **Storytelling:** The ability of the visualizations to tell a story and effectively convey a message to the audience.

- **Integration:** The integration of multiple visualizations and the ability of the participant to effectively use different types of visualizations to tell a complete story.

Machine Learning Award Criteria

These criteria serve as a guideline for the judges to evaluate the submissions and determine the winner of the award for machine learning application. The final decision will be based on a comprehensive assessment of these criteria and the ability of the participant to demonstrate their proficiency and creativity in the field of machine learning.

- **Accuracy:** The accuracy of the machine learning system in categorizing, predicting, or making decisions based on the provided data.

- **Originality:** The originality of the code used in the program and the uniqueness of the approach taken by the participant.

- **Relevance:** The relevance of the machine learning application to the provided data and the problem being solved.

- **Ease of Understanding:** The ability of the participant to effectively communicate and present the results of their machine learning application.

- **Technical Skill:** The demonstration of technical skill and knowledge of machine learning concepts and techniques in the program.

- **Data Analysis:** The ability of the participant to perform a thorough analysis of the provided data and to effectively use the results to inform their machine learning approach.

- **Model Selection:** The selection of the appropriate machine learning model for the problem at hand and the justification for this choice.

- **Model Validation:** The implementation of appropriate validation techniques to ensure the robustness and reliability of the machine learning model.

- **Presentation**: The judges will evaluate the participants' ability to present their research in a clear and organized manner. The presentation should be well-structured, easy to understand, and effectively communicate the key findings and conclusions of the research.

Research Approach Award Criteria

These criteria serve as a guideline for the judges to evaluate the submissions and determine the winner of the award for research approach. The final decision will be based on a comprehensive assessment of these criteria and the ability of the participant to demonstrate their proficiency and creativity. The following are potential criteria that could be used to judge the competition and determine the winner of the award:

- **Relevance of research question:** The judges will evaluate the participants' research questions to ensure they are reflective of the data provided and address a relevant issue or phenomenon. The question should be well thought out, clear, and meaningful.

- **Scientific method:** The judges will assess the participants' use of the scientific method in their research. This includes evaluating the hypothesis, the design of the experiment, the data collection and analysis, and the interpretation of the results.

- **Statistical analysis:** The judges will evaluate the participants' use of statistical analysis or other relevant analytical approaches. The analysis should be appropriate for the data and research question, and the results should be clearly reported and interpreted.

- **Creativity and originality:** The judges will assess the participants' ability to think creatively and come up with original research questions and hypotheses. Participants who demonstrate original thinking and a unique approach to their research will be given higher scores in this category.

- **Interpretation of results:** The judges will evaluate the participants' ability to interpret the results of their research and draw meaningful conclusions based on their findings. The conclusions should be supported by the data and clearly communicated.

- **Quality of presentation:** The judges will evaluate the participants' ability to present their research in a clear and organized manner. The presentation should be well-structured, easy to understand, and effectively communicate the key findings and conclusions of the research.

- **Implications and practical applications:** The judges will assess the participants' ability to identify the implications and practical applications of their research. Participants who demonstrate a clear understanding of the practical applications of their findings will be given higher scores in this category.

Example Approaches to the Dataset

This table summarizes examples of various machine learning, visualization, and research applications for the World Happiness dataset. It provides examples of how this dataset can be used to analyze the relationship between happiness scores and various factors such as income, freedom, family, health, and education. It is important to note that you do not directly use these examples for your project. Instead, we recommend you use these as an inspiration to approach this challenge.

Application Type	Example	Description
Machine Learning	Predicting happiness scores	Use regression or classification algorithms to predict happiness scores based on various factors such as income, freedom, family, health, and life expectancy.
	Dimensionality Reduction	Use dimensionality reduction techniques to identify the most important factors contributing to happiness and well-being.
Visualization	Happiness scores over time	Plot the time series of happiness scores to show changes in happiness over the years.
	Correlation matrix	Plot a correlation matrix to visualize the relationships between different factors contributing to happiness.
Research	Health and happiness	Analyze the relationship between health and happiness by combining the World Happiness dataset with health data.
	Longitudinal analysis	Explore changes in happiness and well-being over time by combining the World Happiness dataset with other longitudinal datasets.