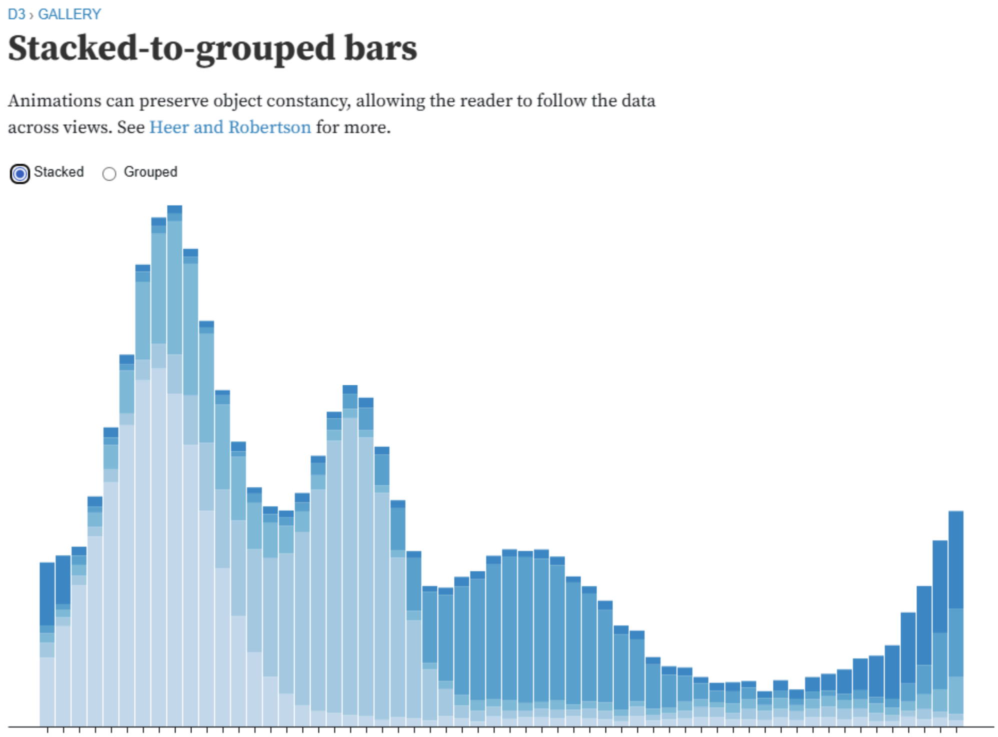
Coby Foy C36763835
Geeth Chilukuri C39295919
Garrett Boling C31811210

Process Book

Overview and Motivation:

- This project's goal is to find out what factors dictate a student's performance in school. We chose this dataset because we wanted to explore a topic that was relevant to us. As college students, we understand how various external factors can affect a student's academic performance. This project gave us the opportunity to further explore this problem domain and utilize thoughtful visualizations to analyze academic performance against varying factors that fight, either positively or negatively, for student success.

Related Work:

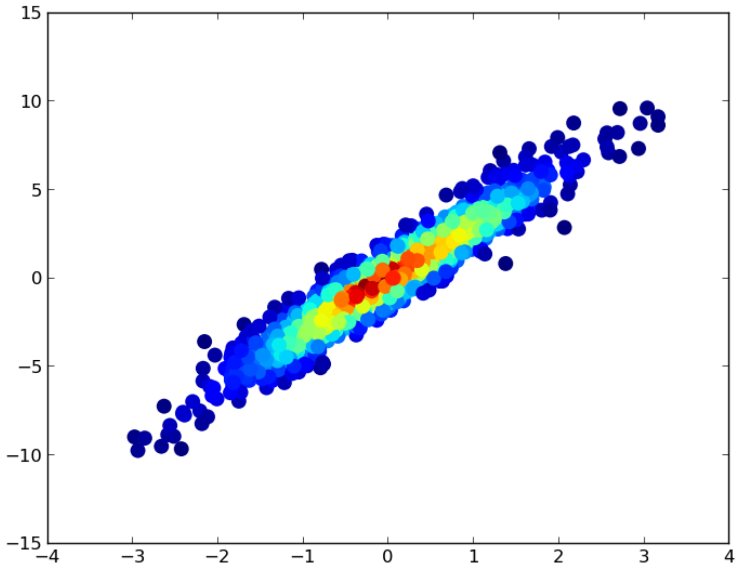


D3 › GALLERY

Stacked-to-grouped bars

Animations can preserve object constancy, allowing the reader to follow the data across views. See Heer and Robertson for more.

- This visualization utilized animations and interactions, "Stacked" and "Grouped" labels to allow the viewer to select their desired layout of the graph. We were inspired to do something similar with one of our graphs where interactions would be well suited since the amount of dots placed on the scatterplot could become overwhelming. Allowing the viewer to filter created a more powerful and interactive visualization.



- We were inspired by this density map which utilizes a diverging color scheme to indicate overall point density. After researching better ways to visualize our data, since a scatter plot displayed points with extremely high density, we decided to visualize something similar to this to show correlation between student performance and various factors that affect it.

Questions:

- What are the highest contributing factors that hold back students?
- What are the highest contributing factors that support student success?
    - The previous two questions have been the driving questions for our project. We aim to determine what factors affect student academic performance.
- Does increased weekday alcohol consumption negatively impact a student's performance on exams?
    - This question emerged as we changed our dataset. The new dataset we analyzed focused on alcohol consumption among students and described how consumption habits affected student's performance on exams. This dataset also described other varying factors which were similar to our previous dataset. We found that increased alcohol consumption negatively impacted student performance.
- What variables contribute to a more "at-risk" student?
    - This question emerged late in our development process as we wanted to utilize aggregate values from a variety of variables to summarize a particular student's "risk profile." We defined a "risk score," that is, a variable that summarizes a particular student's overall likelihood to perform poorly on examinations. The variables taken into account include weekday and weekend alcohol consumption,

    among others. Each variable is weighed to varying degrees, allowing us to favor some factors over others in determining the risk of a student based on circumstance.

- What amount of alcohol consumption is most detrimental to a student's success?
    - This question came to us after discovering that alcohol consumption was a leading factor in lower examination scores within the dataset. We wanted to look further into this by visualizing alcohol consumption against other variables.

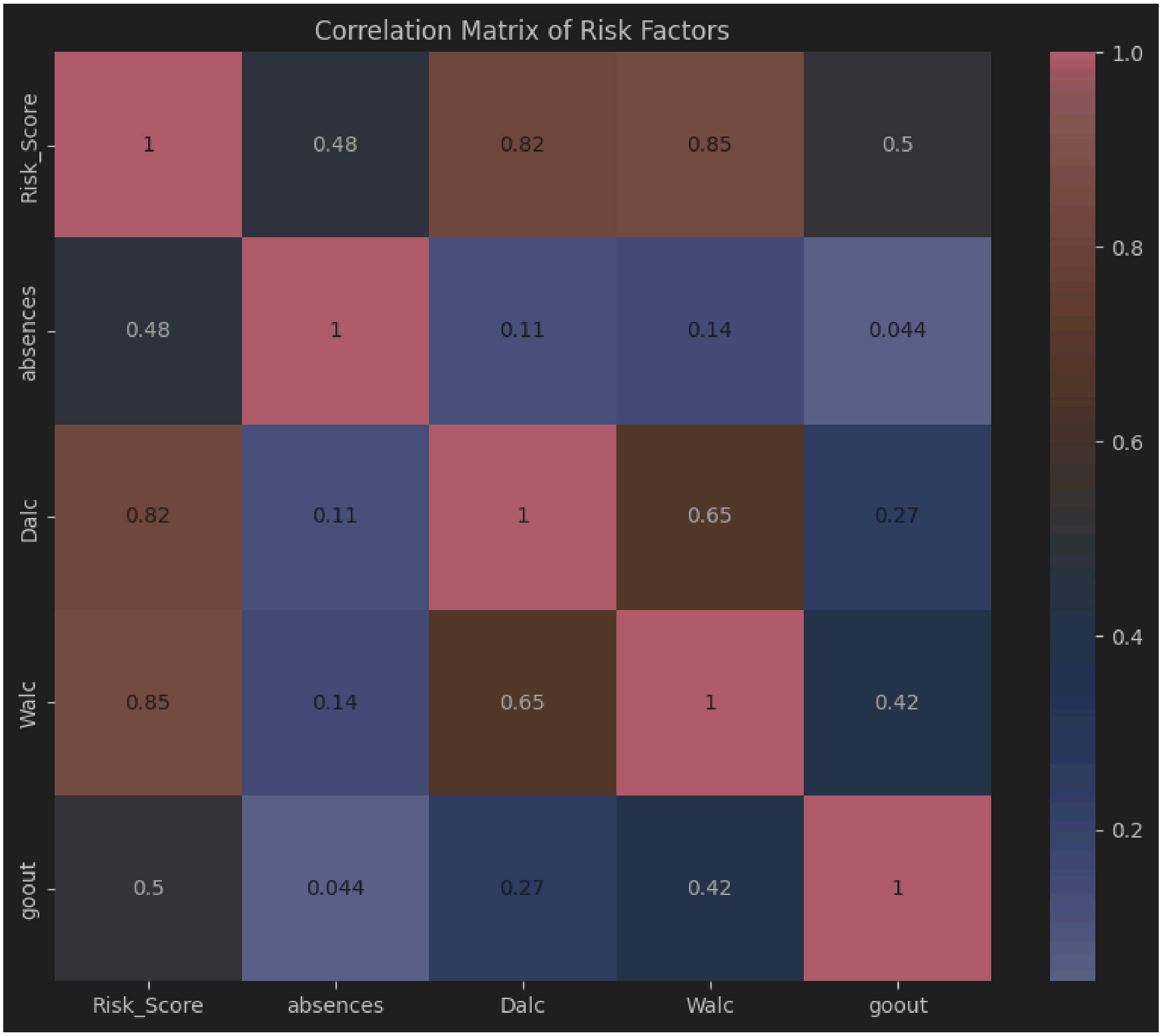
Data:

- We found and downloaded the dataset from kaggle to each of our own personal computers. The kaggle page containing our data contained two datasets, Maths.csv and Portuguese.csv, which we combined to create a dataset with over 1,000 student records. Since each dataset contained common attributes and lacked duplicate entries, combining the datasets was a well-suited choice since both describe various factors affecting student academic performance. The dataset did not contain any null values, so no cleaning was performed.

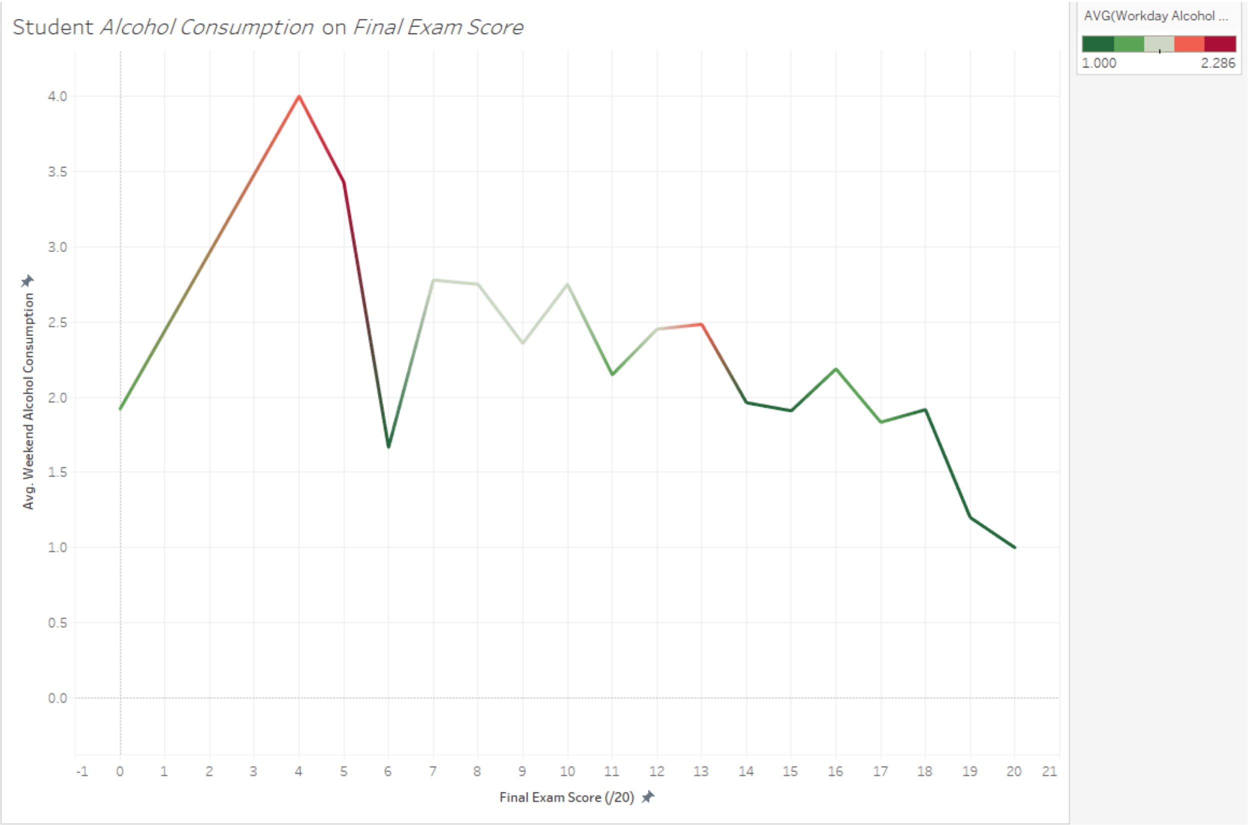
  https://www.kaggle.com/datasets/whenamancodes/alcohol-effects-on-study/data

Exploratory Data Analysis:

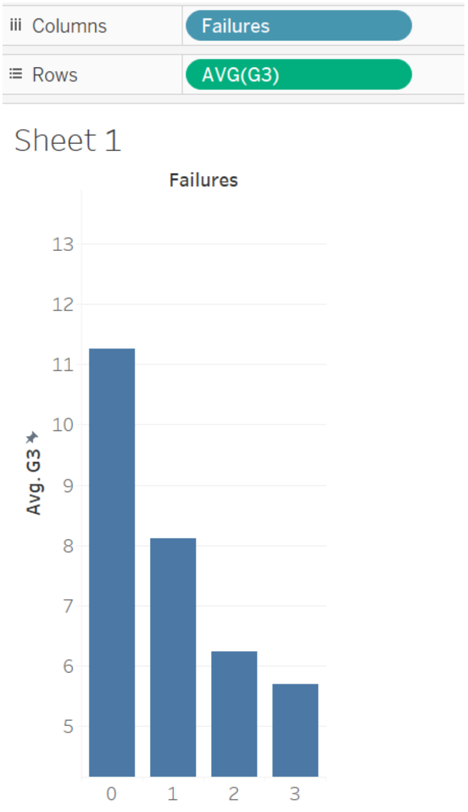
- Initially we explored various variables against student exam scores to determine which attributes tended to support or hurt student academic performance. We initially explored a variety of factors utilizing a correlation matrix to determine what factors appeared to most negatively affect a student's academic performance. From the graph, Dalc and Walc, weekday and weekend alcohol consumption, appeared to have a high correspondence to poor student performance.



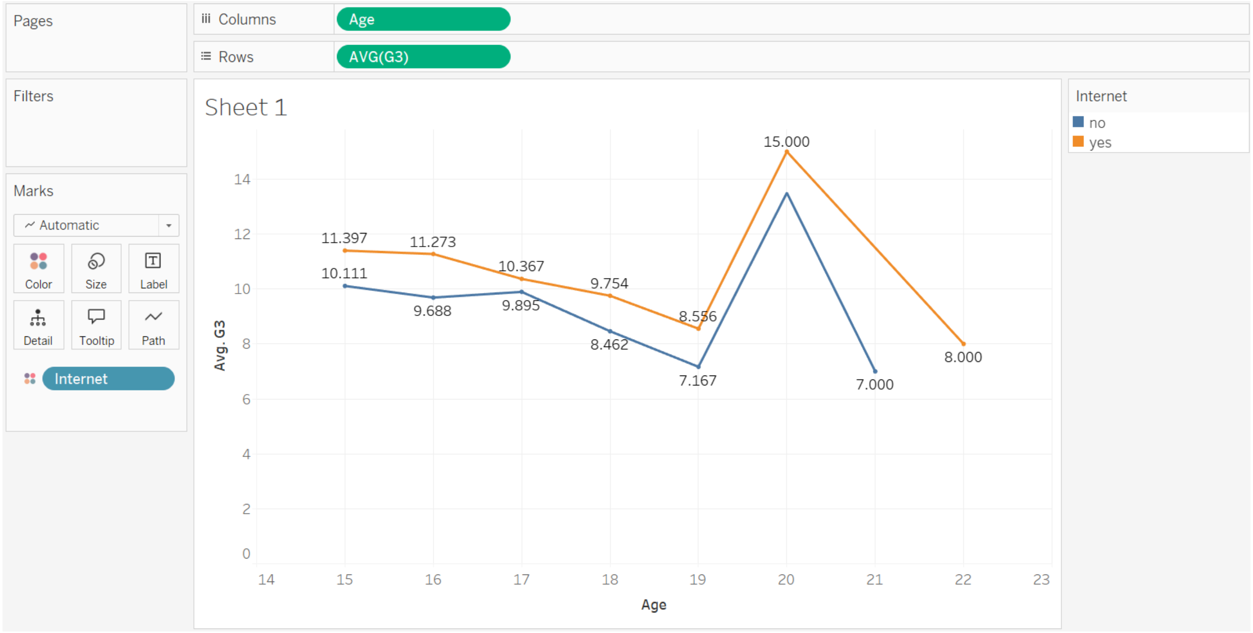
- We began looking at different line graphs and bar charts. We discovered that there was a negative correlation between weekend alcohol consumption and final exam score, where students who consumed more alcohol on the weekend and weekday tended to perform poorer on their final exam.



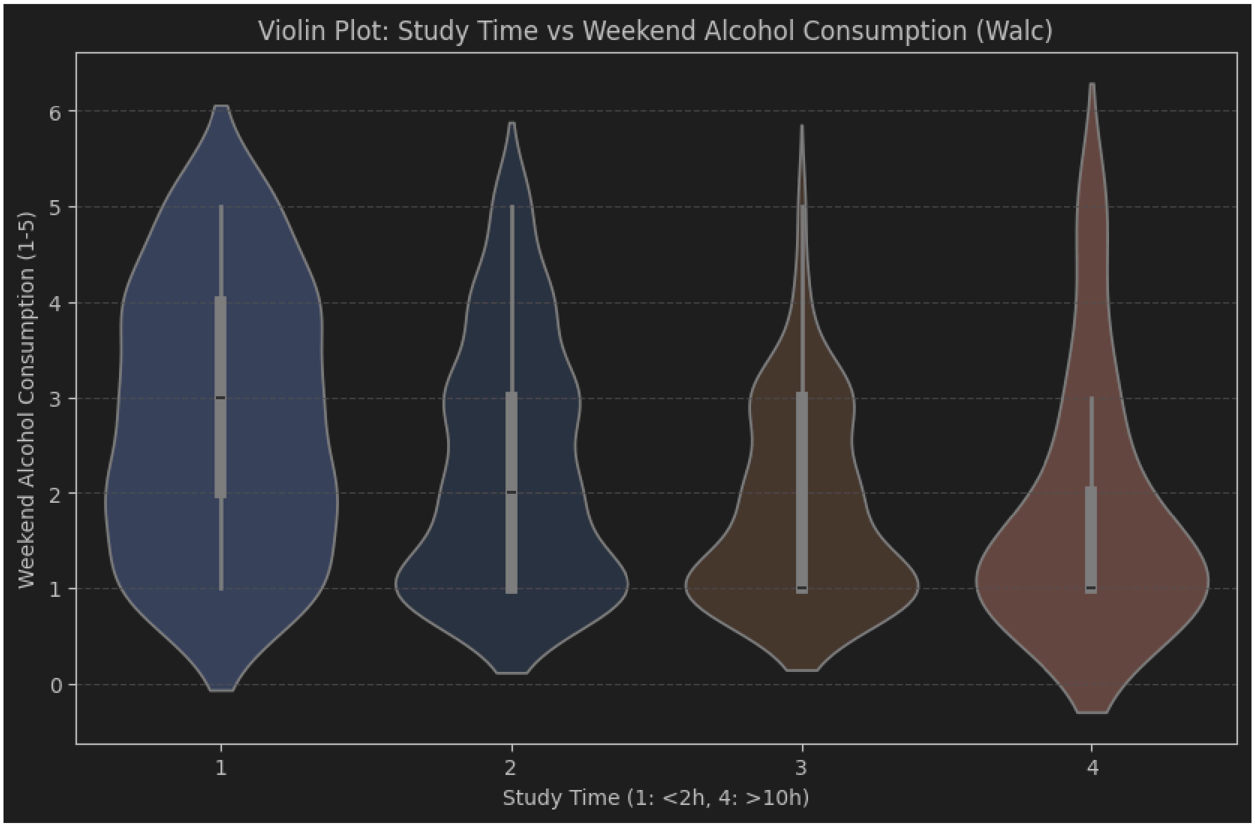
- We also wanted to examine previous class failures as a predictor for exam performance, as visualized by the bar chart below. It appears that students are more likely to perform well on examinations with less previous failures, which intuitively makes sense.



- We also dove into factors such as internet accessibility, discovering that, on average, students with internet access at home performed better academically at every age level than with students who have to travel elsewhere to obtain internet access.



- We utilized a violin plot to analyze the relationship between weekend alcohol consumption and study time. We aimed to determine the correlation between study habits and alcohol consumption.

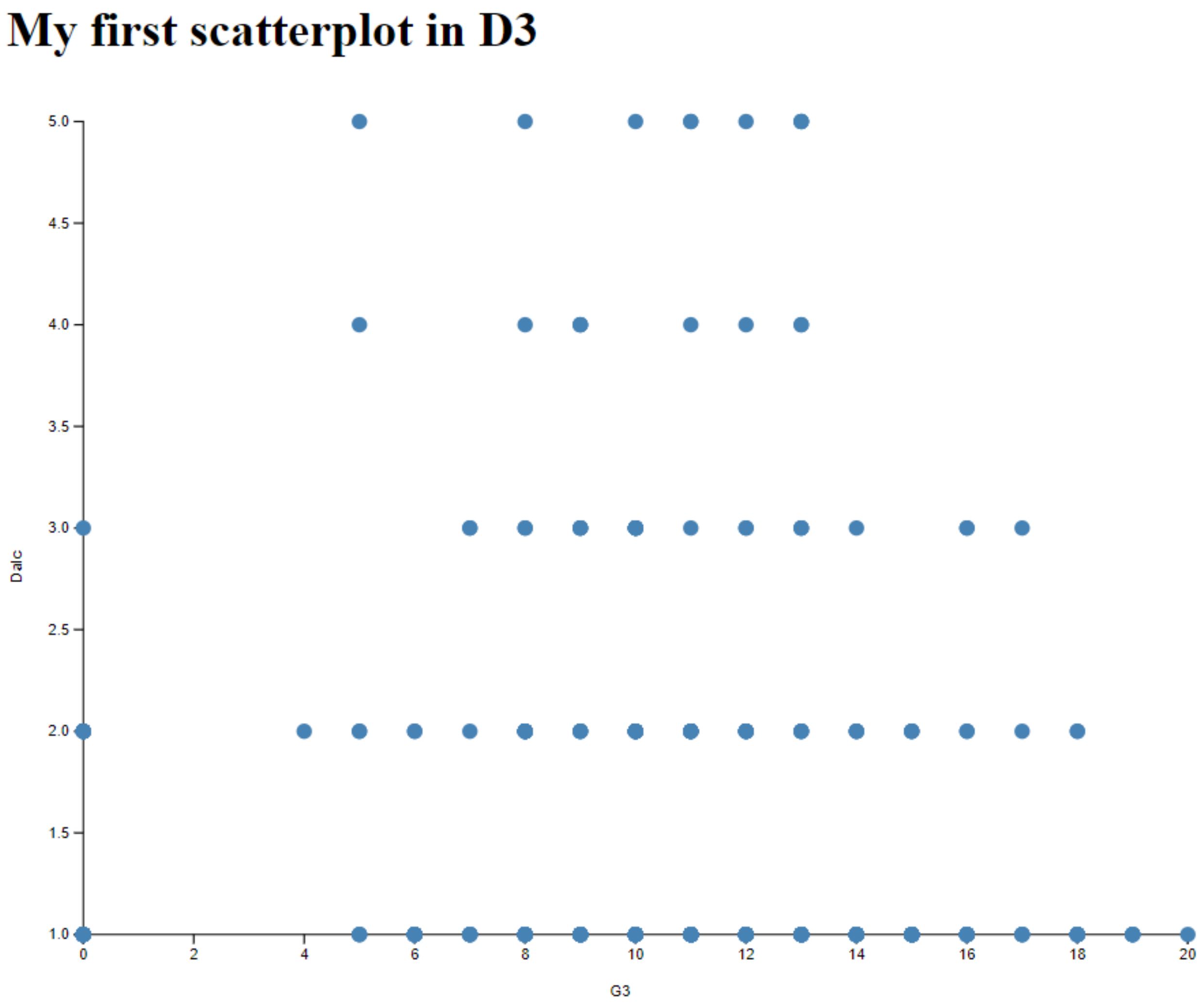


- Graph 1: we first graphed how weekday drinking (Dalc) correlates with final exam scores (G3). We saw that most students were rarely weekday drinkers, and the higher frequency of weekday drinking the lower the average final exam score was.

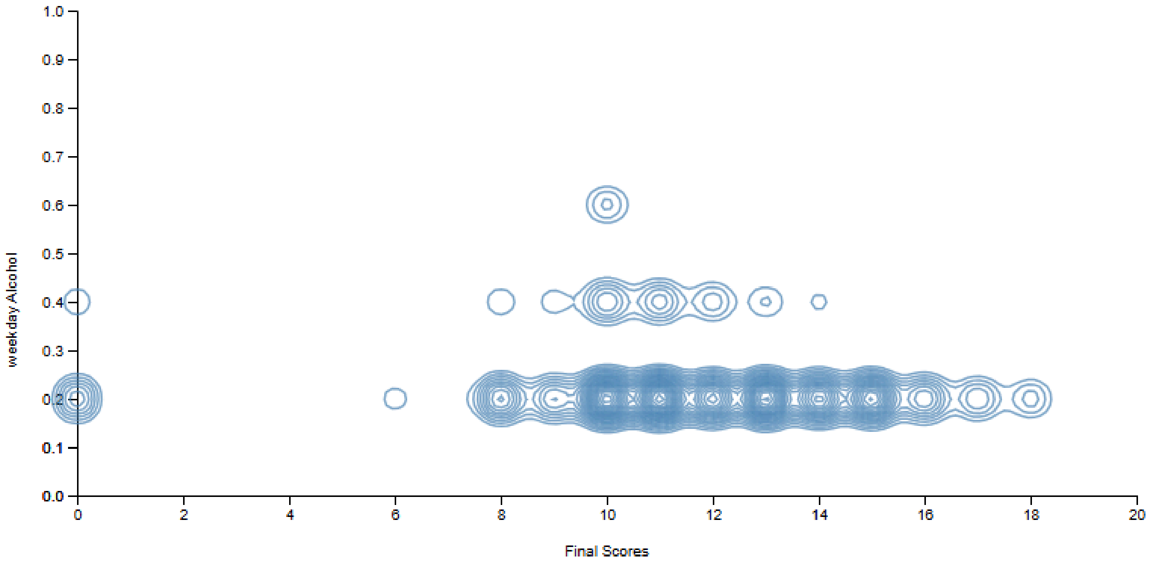
Design Evolution:

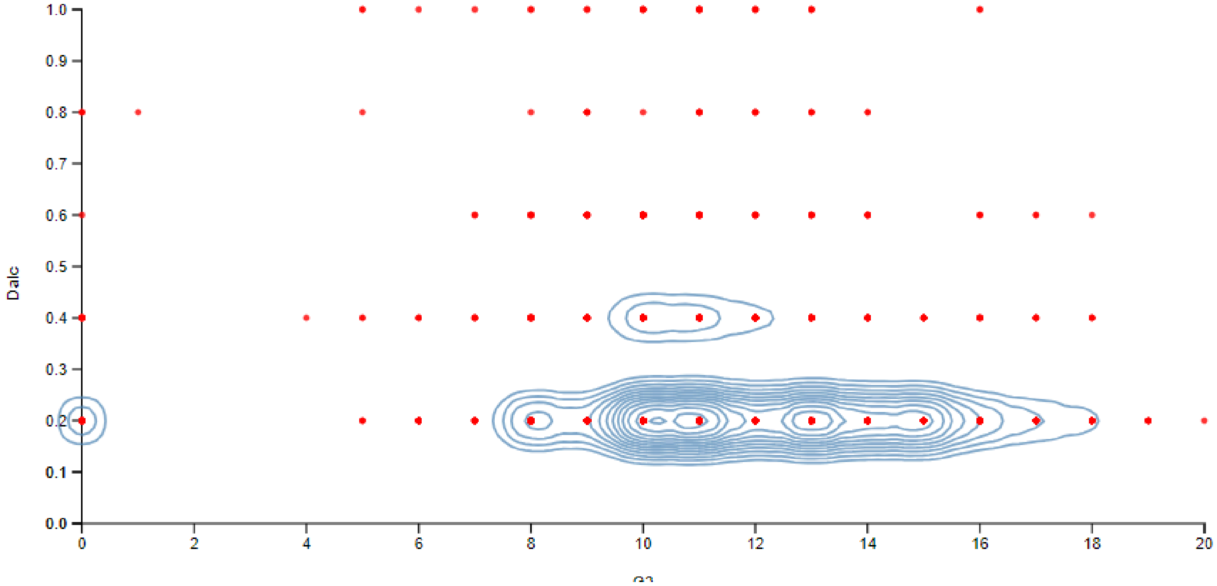
Density Chart

- Graph 1: We first tried to do a simple scatter plot at first but it was hard to visualize density as each point would overlap each other. While it may seem that the scatter plot only has a few handful of points, our dataset contains over 1,000 entries, therefore a significant majority of the points overlap. This made it nearly impossible to distinguish between the individual marks. We needed an alternative that would show the same data but allow for the density of points to be visually represented.

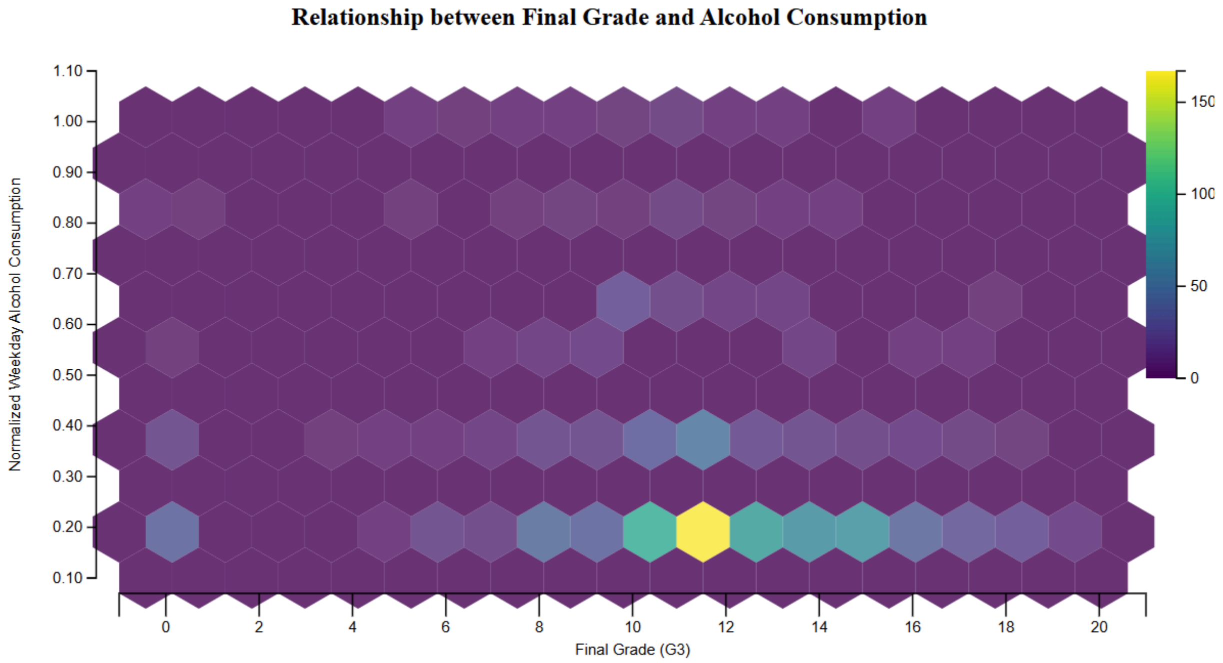


- Graph 2: We decided to utilize a 2-dimensional density chart instead of a scatter plot to show point density. This graph is similar to a scatter plot, but instead of purely individual points, it allows for the true distribution of data over the G3 interval to be portrayed without data points being “hidden” behind others. Overall density of points corresponds to larger circles or clusters of points.



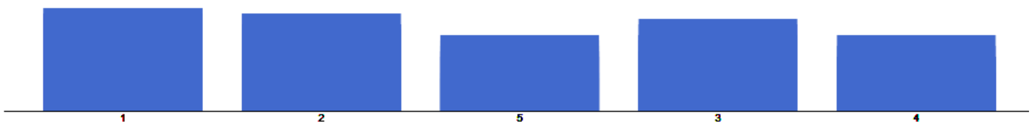


- Graph 3: We decided to utilize a hex-bin density chart instead of a contour density plot to show more of the data visually. This graph is similar to the contour density plot, but instead only showing the places with higher density and having to mouse over the other areas to see the lower frequency, the hex-bin chart is colored to show density.

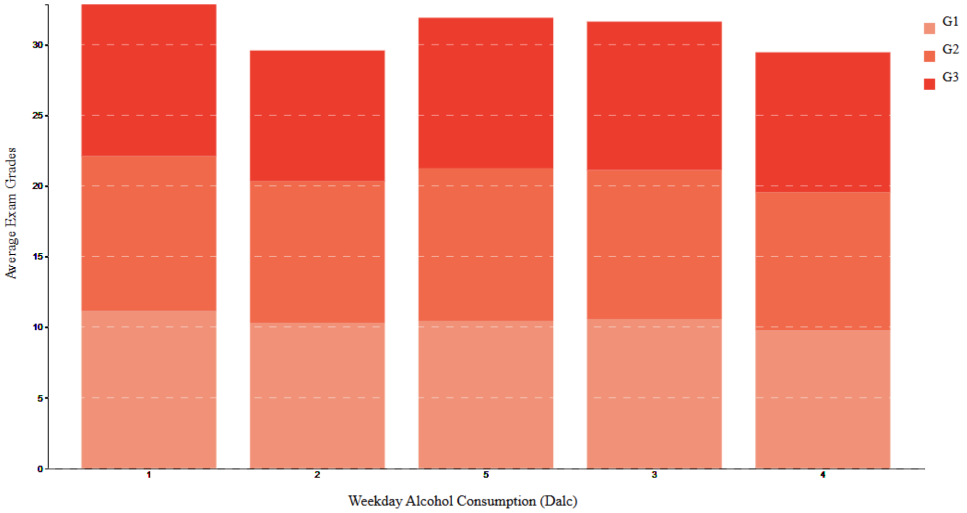


Barchart

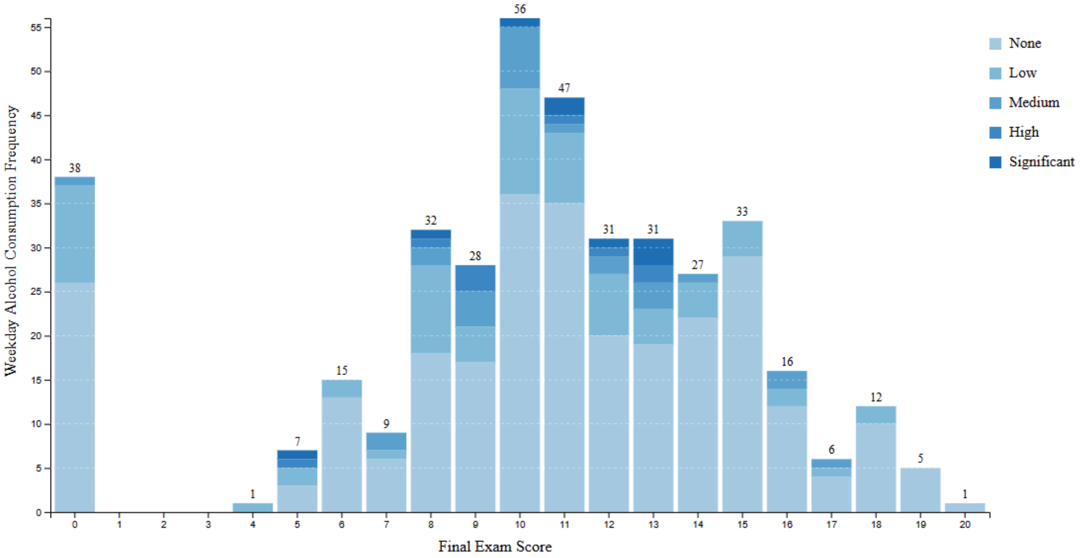
- Graph 1: Our initial graph design choice centered around creating a bar chart where we extracted one variable, G1, representing the first exam score for students. We then created an aggregate function to visualize the total scores for each of the five categories for weekday alcohol consumption (1-5). The scaling is off which contributes to the visualization failing to occupy the space on the graph, making it difficult to see differences between the categories.



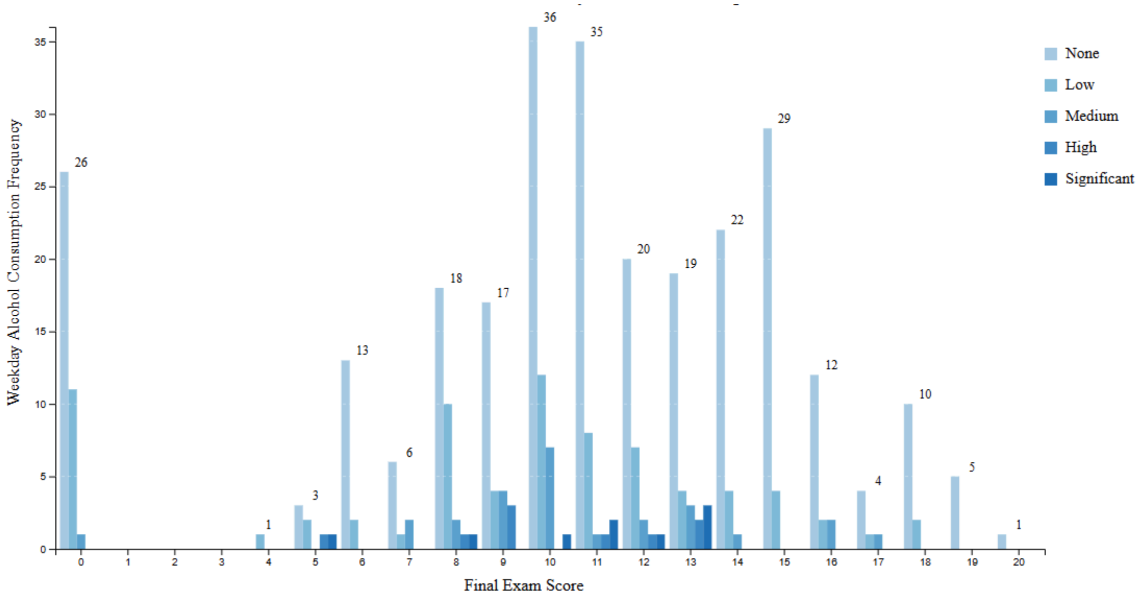
- Graph 2: This graph builds upon the first iteration by adding multiple keys, G1, G2, and G3. These values represent two exams and a final exam score for students. A sequential color scheme is utilized to clearly distinguish between each variable. For each alcohol consumption category, the average exam scores are computed and placed as a stacked bar chart. Reference lines are placed on each tick mark to clearly show the variability among categories of alcohol consumption. Exam score is marked by color type. Average exam score is placed on the y-axis and weekday alcohol consumption is placed on the x-axis into 5 distinct bars. We can see that students who consumed the least amount of alcohol during the week scored higher average scores on all three exams combined.



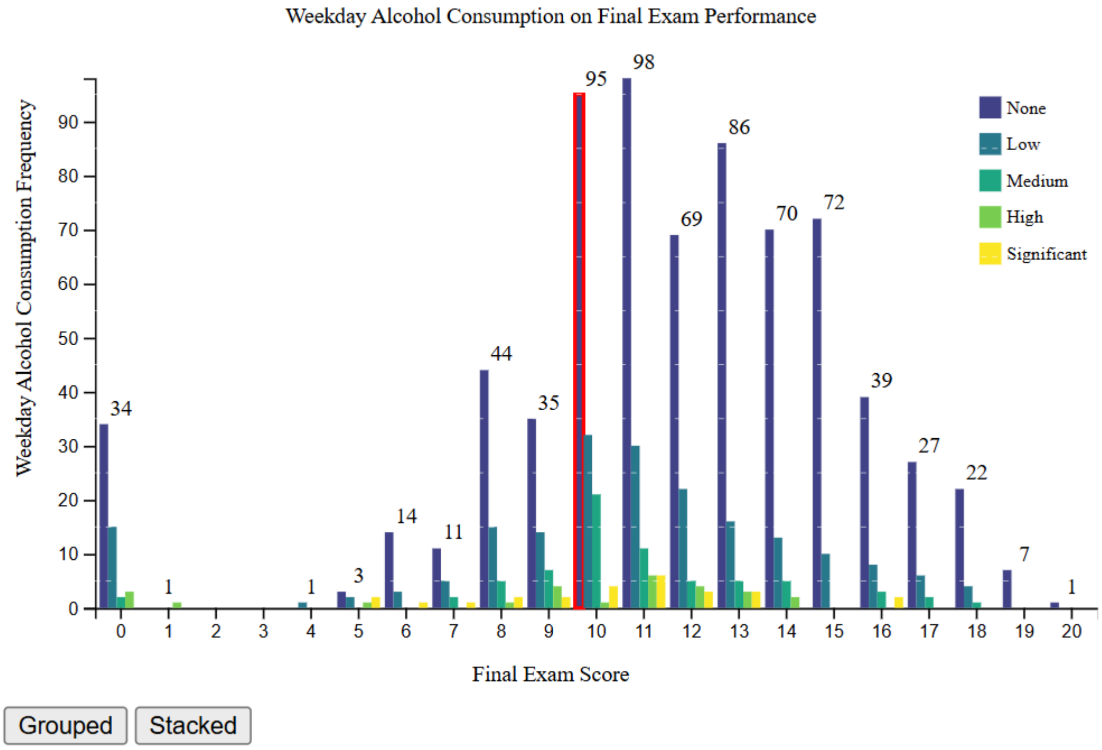
- Graph 3: This barchart iterates off of the previous graph. We noticed that our barchart inhibited our ability to answer the question: *Does increased weekday alcohol consumption negatively impact a student's performance on exams?* The chart fails to show alcohol consumption trends among specific grade distributions, therefore we decided to swap the axes. The result is a barchart that clearly visualizes and answers this question. Now, more details emerge that support our hypothesis, indicating that students who performed very well on exams tend to show lower weekday alcohol consumption patterns. Eureka! The reference lines are still present, but it is difficult to determine quantitatively the amount of alcohol consumption within each exam score category.



Graph 4: A grouped bar chart now allows for a more insightful analysis into the distribution of alcohol consumption within each final exam score group.

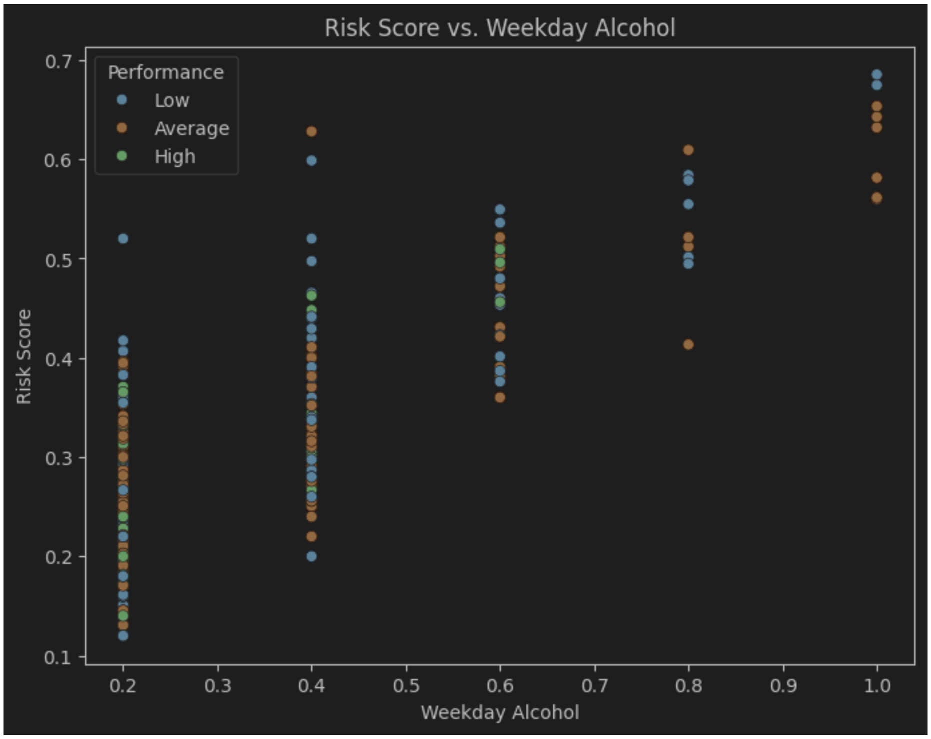


Graph 5: Color scheme changed to more of a qualitative color scheme with higher contrast to allow it to be easier to discern different components. Interactive element added. Upon clicking a given grouped bar, the bar becomes outlined in red and updates values in the scatterplot, see below for more details.

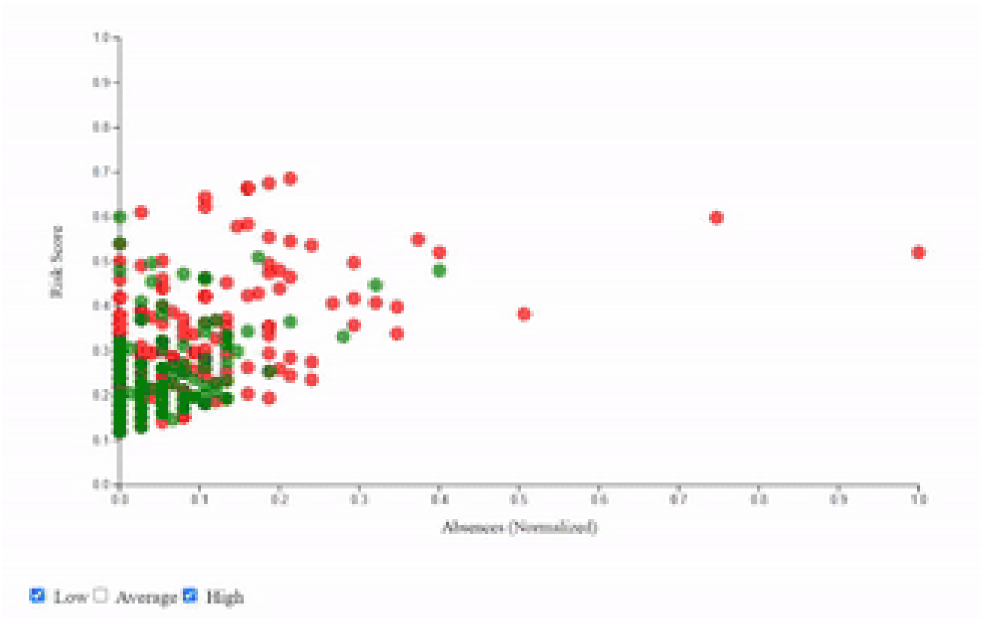


Scatterplot

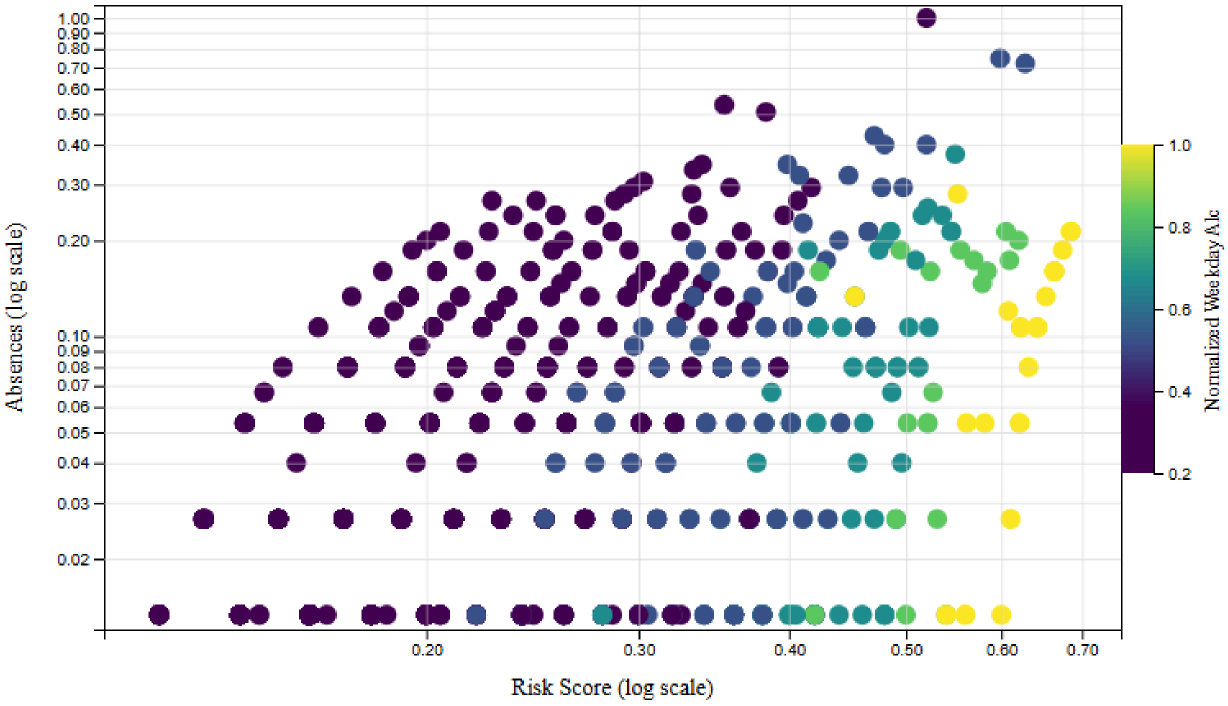
- Graph 1: We started off by throwing our datasets into Python to play around with the data in a familiar environment, which gave us the opportunity to do some feature engineering. We decided to break the student's G3 scores into 3 buckets (low, average, high) to give us a new categorical data type. We also created a new calculated column ("Risk Score") that's based on the number of absences (highest weight), how often they drink during the weekend/on weekdays (weighted lower and higher respectively), and how often they go out with friends (lowest weight). This gave us a good starting point to create some interesting visualizations, where we were able to use the new categorical column to separate out how well students performed using the color channel, and then use the risk score vs. how often students drink during the week to visualize how much "more weekday drinking" actually impacts student's general scores (G3) vs how much it increased their risk score. This is represented in the scatterplot below.



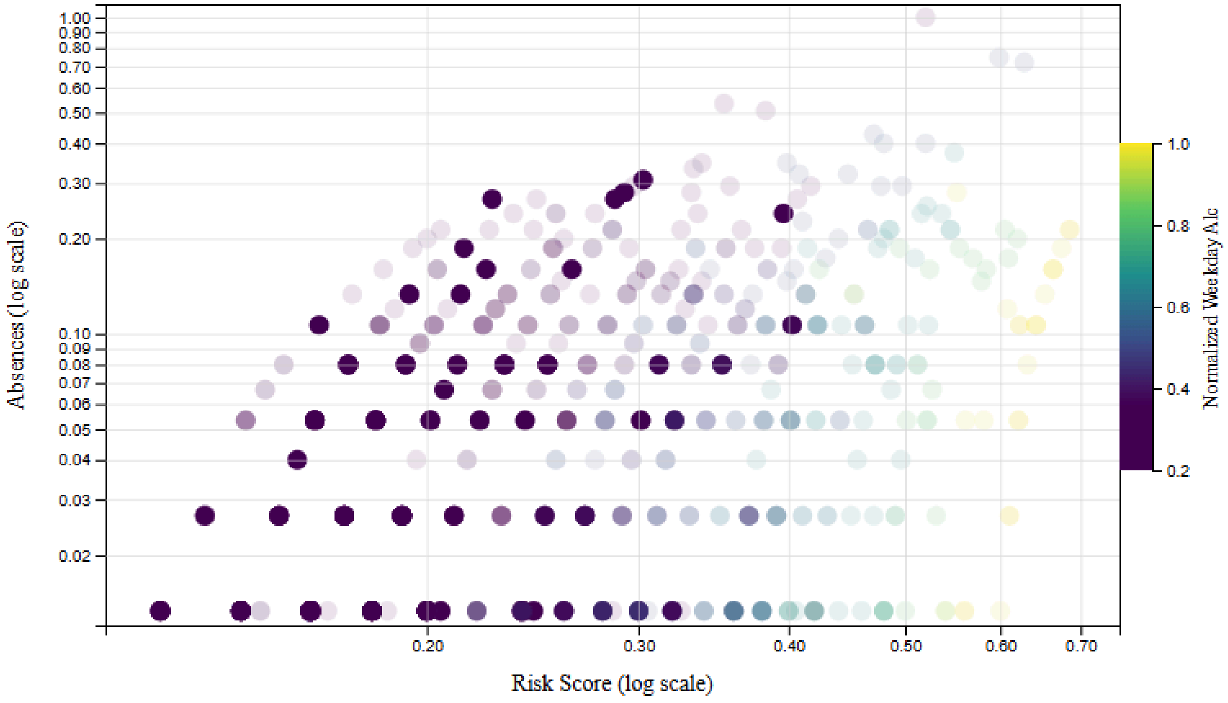
- Graph 2: After doing the initial data exploration, we made the scatterplot easier to read by adding an animation to move the points away from each other on mouseover, showing details about the point that the user is hovering over. We also added a set of checkboxes at the bottom of the graph to allow for an easy way to hide groups based on category. This allows the user to hide the categories if they want to narrow down which group they want to explore further.



Graph 3: After feedback, we decided to use a logarithmic scale to display the results in a more coherent manner, as it keeps the outliers closer to the rest of the data points than the linear scale that we were previously using. We also adjusted the color to represent the normalized amount of alcohol that each student was consuming on weekdays, and switched the axes for readability. After adjustments, the graph shares its color scheme with all other visualizations on the dashboard.



Graph 4: Interactivity added. Once a bar chart group is selected, the corresponding weekday alcohol consumption, Dalc, and final exam score, G3, remain highlighted within the scatterplot. All other remaining dots become less visible, allowing for the viewer to easily discern what student behaviors regarding weekday alcohol consumption on final examination performance correspond to higher risk students. This allows us to more easily identify patterns and helps us answer the question, *What amount of alcohol consumption is most detrimental to a student's success?* Student risk score is a good starting point to discern what students are at-risk for poor academic performance.



Implementation:

- The intent with our interactive elements is to provide the viewer the opportunity to filter the dataset into categories. Since our scatter plot contains a large quantity of markings, we decided to utilize a bar chart filter to allow scatter plot points of the same final exam score and weekday alcohol consumption to remain highlighted while all other points become dimmed.

  Dashboard: https://foycoby.github.io/DataVisualization/
  Repository: *https://github.com/foycoby/DataVisualization*

Evaluation:

- The initial exploratory analysis of our dataset provided us with key indicators for students more likely to be successful and for students less likely to be successful. Since success, in this dataset in particular, is measured by examination scores, we utilized this metric to draw conclusions from our visualizations regarding success. From our initial analysis, we discovered that alcohol consumption, particularly weekday alcohol consumption, correlated with less successful students on average. After our initial question of "What are the highest contributing factors that hold back students" was answered, we decided to utilize the metric "Dalc", or weekday alcohol consumption to further analyze our dataset and produce visualizations that supported our question: "What amount of alcohol consumption is most detrimental to a student's success?" To further explore this, we created a Risk Factor score, which takes into account alcohol consumption and other various factors to predict a student's risk of low academic performance. Through this, our third visualization utilized a risk score to allow us to find an answer to our final question. We determined that, on average, increased weekday alcohol consumption significantly increased a student's risk score, that is, a student's collection of attributes indicating a higher risk of low academic performance.

Project Prototype Milestone:

- What went well: We seemed to produce visualizations with good use of space (data-ink ratio) and good techniques and practices. Our graphs are intuitive and easy to understand. We feel like our visualizations are well constructed and correlate to each other nicely.

- What could be improved: We would like to enhance our visualizations by providing additional interactivity. For example, in the stacked bar chart, we would like to incorporate an interactive element that allows for a viewer to change between stacked and grouped data, similar to what we found in the *Related Work* section. This would allow for each stack to be more easily quantified since all three stacks could be placed side-by-side and better compared and contrasted. This would give the viewer an opportunity to compare average exam scores between and within alcohol consumption categories.

Final Delivery Milestone:

- What went well: Our visualizations became much more coherent and interconnected, providing more of a holistic view of the dataset. Since our analysis focuses mainly on final exam score and weekday alcohol consumption, we decided to incorporate interactivity within our dashboard to further increase pattern recognition, interactivity, and understanding of our data, further allowing us to discern the answer to our initial question of: *does increased weekday alcohol consumption negatively impact a student's performance on exams?* Additionally, our refined graphs appear much more polished, with each graph being upgraded. Our bar chart in particular allows for the user to switch between grouped and stacked–something we wanted to implement after conducting research on what graphs to implement.

- What could be improved: We feel like adding more interactivity could improve the power of our dashboard. Given that we have a few, yet powerful, interactive elements, we feel like adding more could always improve the ability for people to recognize patterns and understand our dataset better with improved interactivity.