Project #4: INTERACTIVE VISUALIZATION USING JAVASCRIPT, R, or PYTHON

EN 605.662.SU20 Data Visualizations

Johns Hopkins University

11/29/2020

Abstract

Data Scientists have a wide range of tools at their disposal for data visualization purposes. This project sought to expose students to some of the popular libraries available in JavaScript, R, and/or Python. Three separate data sets were utilized to offer a wide variety of visualization styles.

Description of Project 1: Student Performances

This project sought to explore the influence of multiple variables on student test performance. The data set includes scores from three standard student exams (math, reading, and writing), as well as a variety of student and parental socioeconomic factors that interact with student performance. By identifying key factors and interactions between these variables, we might better understand the effectiveness of student test preparation courses, the strongest factors which contribute to test outcomes, the impact of parental education levels, and how best to improve student scores on each test type.

Data Set

The data set utilized, Students Performance in Exams, was obtained from Kaggle (https://www.kaggle.com/spscientist/students-performance-in-exams/discussion/160544). There are 1,000 observations in the data set and it includes the following 8 variables: Gender, Race/Ethnicity, Parental Level of Education, Lunch, Test Preparation Course, Math Score, Reading Score, and Writing Score. Gender, Race/Ethnicity, Parental Level of Education, Lunch, and Test Preparation Course are all categorical variables (presented as character data types). Gender was classified as either male or female. Race/Ethnicity was classified as: group A, group B, group C, or group D. There were no descriptions as to what these groups meant. It was not explicitly stated whether this was parental or student race, but it was assumed the two were equivalent. Parental Level of Education was classified as: some high school, high school, some college, Associates degree, Bachelor's degree, or Masters degree. Lunch was classified as standard or reduced. This variable indicated whether the student received a discounted or free lunch at school. Test Preparation Course was classified as either none or completed and indicated whether the student had completed a class designed to improve test outcomes.

The data set utilized, Students Performance in Exams, was obtained from Kaggle (https://www.kaggle.com/spscientist/students-performance-in-exams/discussion/160544). There are 1,000 observations in the data set and it includes the following 8 variables: Gender, Race/Ethnicity, Parental Level of Education, Lunch, Test Preparation Course, Math Score, Reading Score, and Writing Score. Gender, Race/Ethnicity, Parental Level of Education, Lunch, and Test Preparation Course are all categorical variables (presented as character data types). Gender was classified as either male or female. Race/Ethnicity was classified as: group A, group B, group C, or group D. There were no descriptions as to what these groups meant. It was not explicitly stated whether this was parental or student race, but it was assumed the two were equivalent. Parental Level of Education was classified as: some high school, high school, some college, Associates degree, Bachelor's degree, or Masters degree. Lunch was classified as standard or reduced. This variable indicated whether the student received a discounted or free lunch at school. Test Preparation Course was classified as either none or completed and indicated whether the student had completed a class designed to improve test outcomes.

Using R, descriptive statistics were conducted on all variables in order to obtain their frequency, mean, median, mode, minimum, maximum, ranges, standard deviation, variance, skew, and kurtosis values. Exploratory data analysis utilized scatterplots, histograms, and boxplots for individual variables to better understand their central tendencies, variability, and to detect outliers.

Though not for the purposes of the required visualizations of this assignment, the following were created for exploratory purposes: Test score boxplots and histograms were created in plotly for each test score (math score, reading score, writing score); grouped boxplot comparing the count (or frequency) or each test score (math, reading, writing); stacked vertical

line and bar graphs were created to evaluate gender differences on student performances on each test score (math, reading, and writing); horizontal and vertical boxplots were created to evaluate the effect of lunch status on student performance on each test score (math, reading, and writing); side by side grouped vertical bar charts were created comparing gender and test grades (A, B, C, D) for each test type (math, reading, writing); and overlapping histograms for each test score by gender.

Visualizations and Updates

Grouped Vertical Dot Plot. To explore the impact of parental upbringing on student test performance, vertical dot plot were used to compare Parental Level of Education, Race/Ethnicity, and test score (math, reading, and writing; see Figure 1.). The vertical dot plot was selected because of the large amount of information presented for this analysis (both in terms of total data points and within each of the two categorical variables). Rather than using multiple vertical bar charts, a vertical dot plot could more concisely present the information and allows for better and more accurate interpretations due to the interactivity components of Plotly. The dot plot also allows users to identify both gaps and clusters in the visualization, as well as an opportunity to see how the data spreads along the y-axis due to the improved data-ink ratio. Although overplotting can be a problem with grouped dot plots (a problem avoided by grouped bar charts), this issue is addressed by including user-interactivity and/or filter/selection functionality into the visualization.

For each test score, vertical dot plots displayed Parental Level of Education on the x-axis and test score on the y-axis. Each data point was color-coded to categorize it by Race/Ethnicity. For clarify, a legend was provided. To allow for user-interactivity, Plotly was utilized and tooltip labels were modified to include specific information for each data point. Users could double click on an individual Race/Ethnicity legend to isolate it's values across each Parental Level of Education, lasso select specific regions, zoom in on specific regions, or hover their curser over a specific data point. Additionally, a filter event was created to remove existing marks and rescale axes to a user-defined grouping variable. As seen

in Figure 2, when added to the Plotly visualization, users could selectively filter for Race/Ethnicity, such that only the selected group is displayed. Although this seemed to effect the visibility of the x-axis tick labels and the jitter of each data point, it did greatly improve readability of each category. This feature does appear redundant, as the same function is also included as part of Plotly's visualization.

Overall, it appears that most Parental Levels of Education yield relatively similar distributions on math scores. However, there does appear to be greater variability and/or more outliers in the "some high school" group, and many fewer data points in the "master's degree" group. Most Race/Ethnicity groups had relatively similar distributions as well. However, group E did appear to have a slightly higher minimum value on math scores. The distribution of math test scores appears relatively normally distributed across each parent category, with a higher density of points aggregated around 75, and fewer points at the upper quadrant (100) and lower quadrants (25). A vertical bar chart would be useful to confirm.

Although not discussed here, the same visualizations were created for all three test scores.

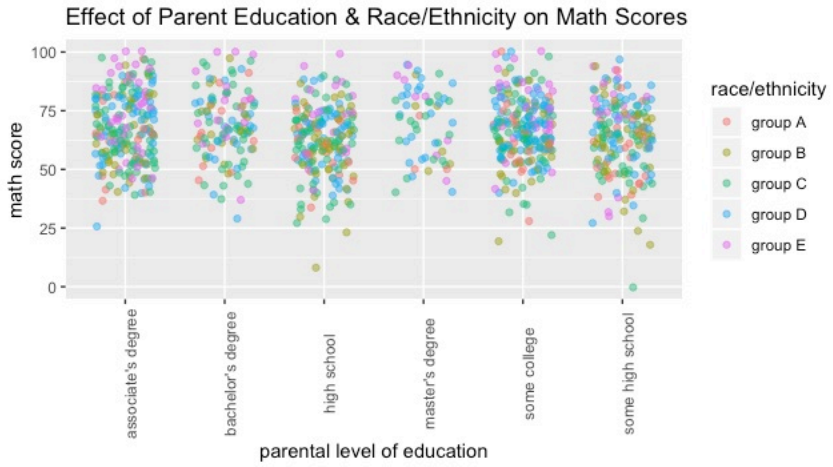

FIGURE 1. Vertical dotplot displaying effect of Parental Level of Education on Math Score, Grouped by Race/Ethnicity

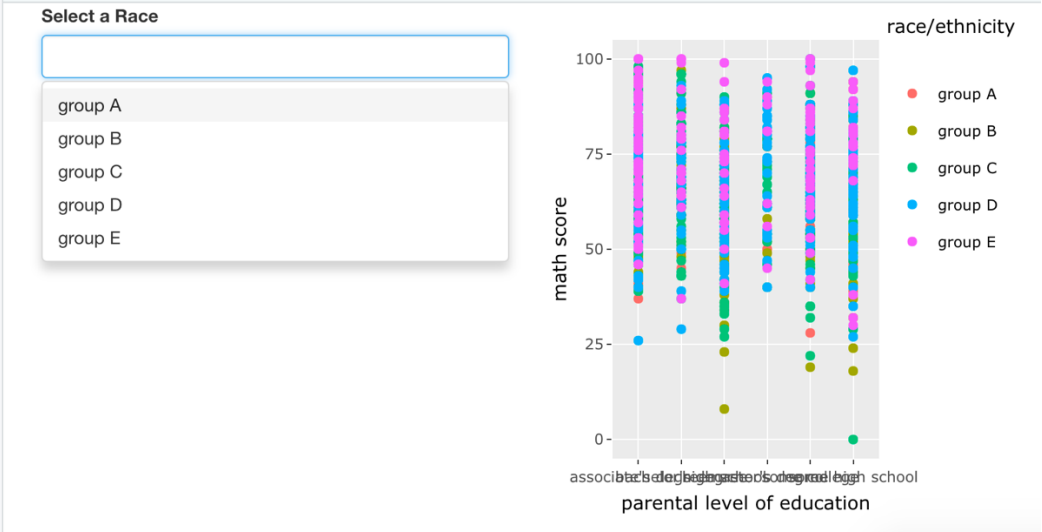

FIGURE 2. Vertical dot plot displaying the effect Parental Level of Education and Race/Ethnicity on student math scores.

Grouped Vertical Bar Char. The previous visualization examined the effect of parental variables on student test performance. To begin examining student-related variables, a grouped vertical bar chart was created to display the frequency of math test scores categorized into "grade bins" and grouped by gender (see Figure 3). Bar charts are useful visualizations when representing data gathered into discrete value categories. Each bar is grouped by these discrete categorical variables and is plotted against a numerical value on the opposing axis. The length or height of the bar dictates the value of the category. This type of chart affords quick and accurate information decoding due to the fact that human perception is much more reliable when judging length (or, position along a scale) than it is when judging areas (or, angles). This easy information decoding means that grouped vertical bar charts are useful when representing distributions of data points or comparison of data points across different subgroups of data. Viewers can quickly ascertain highest or most common categories or determine similarities or differences between them.

Here, each test score was discretized into bins, according to standard grade categories (A=90-100%; B=80-89%; C=70-79%, D=60-69%, F=0-59%) and displayed on the x-axis. The frequency of occurrences in each bin (or, count) was represented on the y-axis. Discrete gender categories were represented as different color boxes and were visualized side by side for each grade bin. This process was repeated for each subject (math, reading, writing). Although test scores themselves represent a continuous variable, grades represent a discrete category. As such, the decision was made to represent this relationship as a grouped vertical bar chart rather than a grouped (or, overlapping) histogram (although, this is provided for comparison; see Figure 4).

The grouped vertical bar chart was created using Plotly, to improve user-interactivity. In this way, users have the option to trace data points, isolate categories, or select ranges of values. A chart title was added for clarity. And, a user-defined filter drop-down menu was also included as a means to selectively view each gender. Although some users may find the drop-down filter menu more intuitive, its feature is redundant given the functionality of Plotly.

Upon initial inspection of the grouped vertical bar chart, it is apparent that the most frequent math test grade is “C”, followed by “B”, “D”, “A”, and then “F”. This pattern holds for both male and female students, although the difference between “B” and “C” for each gender is significant. Overall, the frequency of values appears like a normal distribution (though, this is not a histogram the y-axis is representing frequency values).

Though not presented below, this analysis was repeated for each subject (images included in screenshots folder submission). Interestingly, math grade distributions differ slightly from both reading and writing grade distributions in both males and females. Future visualizations may opt to include all test subjects by age and gender.

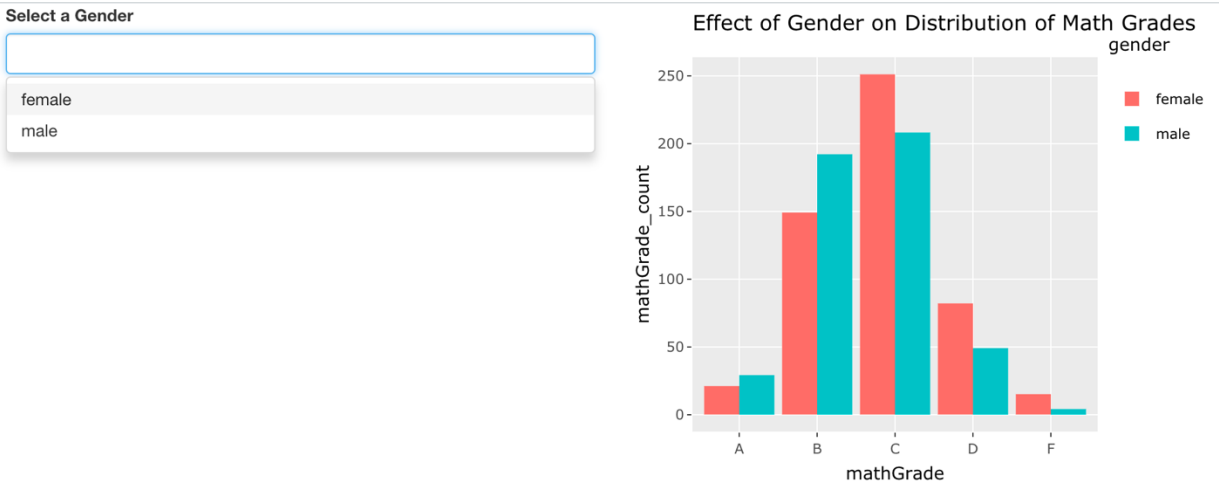

FIGURE 3. Grouped vertical bar chart exploring the relationship between frequency of math grades and gender.

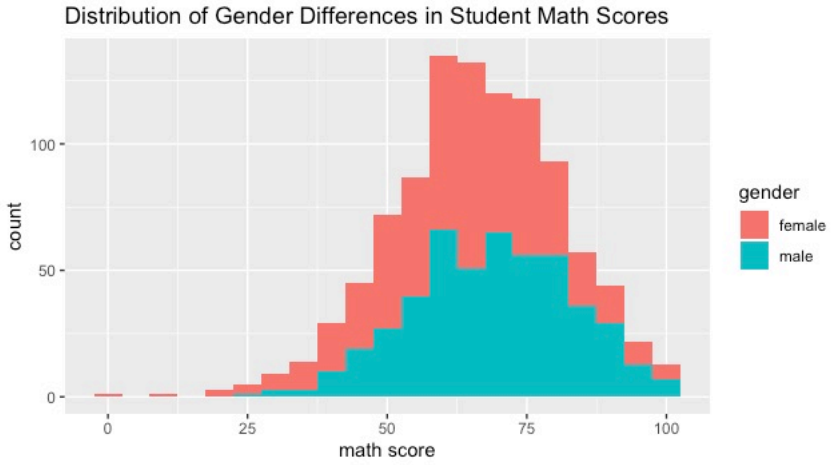

FIGURE 4. Overlapping histograms for math score frequencies grouped by gender.

Correlation HeatMap. Since most visualizations utilized showed a relatively similar distribution pattern for each test subject, a correlation heatmap was created (see Figure 5). Correlations demonstrate the strength of a relationship between two variables. A correlation coefficient value ranges anywhere from -1 to +1, where -1 indicates a strong negative relationship, +1 indicates a strong positive relationship, and 0 indicates no relationship at all. More specifically, correlation coefficients indicate that for every positive or negative increase in one variable, there is a similar increase of a fixed proportion in the other variable.

To visualize a correlation matrix, a correlation heatmap uses a grid of colored cells, filled along a continuous color scale depending on the value of the correlation between two discrete variables. The grid acts as coordinates between two discrete joint distributions. Each cell has an equal width. In a typical heatmap, the entire cell is filled by color. Thus, color strength is the only factor indicating the strength of a relationship (i.e., size and shape are held constant). Due to the fact that color is proportional to the value of the correlation, viewers can quickly identify incidence patterns. The stronger the color (i.e., stronger in shade, color, or brightness), the stronger (or larger) the magnitude of correlation.

Here, the correlation heatmap was created with Plotly, which uses a user-driven trace function, allowing users to interact with each cell block. Each cell provides a tooltip labeling the two variables and their correlation coefficient. For clarity, a title was added to the heatmap and a color legend was provided. Due to the small number of numerical variables compared, the heatmap is relatively easy to understand. However, the colors are all presented in a single color scale, which can make it difficult for the eye to focus on any one relationship in particular. This is a limitation of most correlation heatmaps, which suggests that it may be beneficial for future presentations to include a size parameter to the heatmap, such that the size of each square corresponds to the magnitude of the correlation it represents.

We see that all tests are, indeed, correlated with one another. All squares are filled in blue, representing positive correlations for each variable. We see that writing and reading scores that the strongest correlation ($z=0.95$), followed by reading and math ($z=0.82$), and finally writing and math ($z=0.80$). It should be noted, however, that all of these correlation coefficients still indicate strong relationships between the three variables, such that an increase in one indicates a corresponding increase in the other.

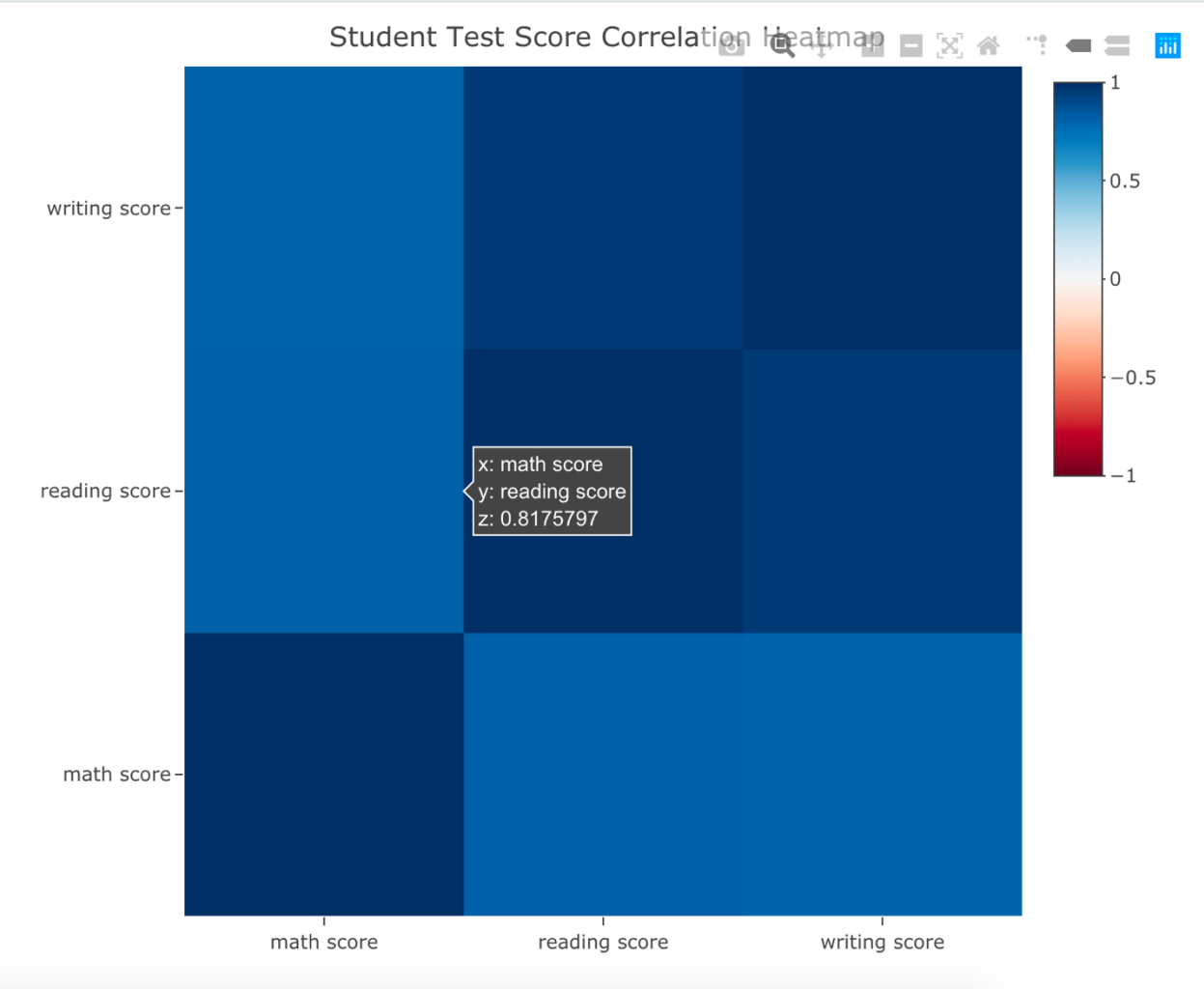

FIGURE 5. Student test score correlation heat map. Compares correlation coefficient for each test score (math, reading, and writing).

Description of Project 2: COVID-19 in South Korea

This project sought to examine confirmed cases of Coronavirus-19 (COVID-19) in South Korea. To date, South Korea has 34,652 confirmed cases, of which 27885 have recovered and 526 have died. The data set, provided by the Korea Centers for Disease Control and Prevention, tracks the number of confirmed cases by day between January 20th, 2020 and March 9th, 2020, as

well as a variety of other patient-specific factors. By identifying key factors and interactions between these variables, we might better understand the progression and spread of this novel disease.

Data Set

The data set, Data Science for COVID-19, was obtained from Kaggle (https://www.kaggle.com/kimjihoo/coronavirusdataset?select=Case.csv). To limit analyses to patient-specific behaviors and demographics, only the PatientInfo.csv was utilized. This data set included 7754 observations and the following 18 variables: global ID, local ID, sex, birth year, country, province, disease, group, exposure start date, exposure end date, infection reason, infection order, infected by, contact number, confirmed date, released date, and deceased date. Global ID is a numerical variable used as a unique patient identifier and represented as a double datatype. Sex is a categorical variable with two levels (male or female) and is represented as a character datatype. Birth year is a numerical variable used to identify the age of the patient and is represented as a double datatype in the format two-digit month/two-digit day/two-digit year. Country is a categorical variable indicating a patient's country of residence and is represented as a character datatype. Province is a categorical variable indicating a patient's province of residence and is represented as a character datatype. Disease is a numerical variable represented as a double datatype.

Group is a binary categorical variable indicating represented as 1 if the patient was part of a group infection and 0 otherwise. Exposure Start is a numerical variable used to identify the date the patient was initially exposed to COVID-19 and is represented as a double datatype in the format two-digit month/two-digit day/two-digit year. Exposure End is a numerical variable used to identify the last date the patient was exposed to COVID-19 and is represented as double

datatype in the format two-digit month/two-digit day/two-digit year. Taken together, the difference between Exposure End and Exposure Start dates indicates the length of exposure for each patient. Infection Reason is a categorical variable indicating the way in which the patient was exposed to COVID-19 and the variable is represented as a character datatype. Infection Order is a numerical variable represented as a double datatype, though it is unclear what this variable represented.

Infected By is a numerical variable representing the patient ID of the individual who infected the current patient. Contact Number is a numerical variable indicating the number of people with whom the patient had contact, and is represented as a double datatype. Confirmed Date is a numerical variable indicating the running total number of confirmed COVID-19 positive patients on the specified date and is represented as a double datatype in the format two-digit month/two-digit day/two-digit year. Released Date is a numerical variable indicating the running total number of recovered COVID-19 patients on the specified date and is represented as a double datatype in the format two-digit month/two-digit day/two-digit year. Finally, Deceased Date is a numerical variable indicating the running total number of deceased COVID-19 patients on the specified date and is represented as a double datatype in the format two-digit month/two-digit day/two-digit year.

Using R, descriptive statistics were conducted on all variables in order to obtain their frequency, mean, median, mode, minimum, maximum, ranges, standard deviation, variance, skew, and kurtosis values. Exploratory data analysis utilized scatterplots, histograms, and boxplots for individual variables to better understand their central tendencies, variability, and to detect outliers. Exploratory data analyses were conducted in R's ggplot and Plotly.

Though not for the purposes of the required visualizations of this assignment, the following were created for exploratory purposes: age histogram, age group vertical bar chart, age group by gender stacked vertical bar chart, province vertical bar chart, patient infection reasons by province vertical bar chart, age by province heat map, age by province by gender heat map, infection network analysis, and age by province and gender stacked vertical bar chart. Code and screenshots for these accessory visualizations are provided in the .zip file.

Visualizations and Updates

Grouped Vertical Bar Chart. To examine age and gender differences in frequency of confirmed cases across each province, a grouped vertical bar chart was created using Plotly (see Figure 6). As mentioned earlier, bar charts are useful visualizations when representing data gathered into discrete value categories. The length or height of the bar dictates the value of the category, affording quick and accurate information decoding. When grouped across different subgroups, viewers can easily compare heights, common categories, similarities, and differences.

Here, age was discretized into 10-year bins and displayed on the x-axis. The frequency of occurrences in each bin (or, count) was represented on the y-axis. Discrete gender categories were represented as different color boxes and were visualized side by side for each age bin. This process was repeated for each province. Although age can be represented a continuous variable, it's easier to draw comparisons between groups when age is discretized. As such, the decision was made to represent this relationship as a grouped vertical bar chart rather than a grouped (or, overlapping) histogram.

Since there were 12 provinces in total, an overlapping or grouped chart would have been entirely too difficult to read. As a result, each province was represented in its own grouped vertical bar chart and was presented using a facet wrap. To make it easier to isolate a specific

province, a check box filter was provided, allowing users to isolate one province, compare multiple provinces, or examine all provinces at once. The plotly function allows users to double-click on a specific gender legend to isolate gender across each province chart.

Here we see the higher number of confirmed cases in the capital area, followed by: Gyeongsangbuk-do, Daegu, Daejeon, Gwangju, filtered at airport and Jeju-do, Jeollabuk-do and Jeollanam-do, Ulsan and Chungcheongbuk-do, and Busan. It's possible these differences are due to population size or population spread differences and future analyses should examine both factors for each province. We also see that the capital area appears to have more male confirmed cases compared to all other provinces. It's unclear whether the capital area or Gyeongsangbuk-do have more female confirmed cases. Busan is the only province with male-only confirmed cases, although there was only one confirmed case. Gangwon-do was the only province with female-only confirmed cases, although there was only one confirmed case. Very few provinces had confirmed cases in patients aged 0 to 9 (only observed in Gyeongsangbuk-do and the capital area). Similarly, few provinces had confirmed cases in patients aged 90-99 (only observed in Gyeongsangbuk-do and Daegu). Future analyses may opt to include life expectancy to determine whether these observations are normal.

Within the capital area, there are significant gender differences across age bins. Males follow a relatively normal distribution, with higher frequencies of cases between the ages of 30 and 59 (maximum between 30 to 39). Male confirmed cases drop both above and below these bins. Females, however, have the highest frequency of confirmed cases between 20 and 29, which slowly declines as age bins increase.

Within Gyeongsangbuk-do, it appears there are more female confirmed cases in most age bins (two exceptions: 20-29 and 80-89). Both genders have an abnormal distribution with higher

frequencies between 20 to 29, a decrease in frequencies until increasing again at 50 to 59 (female) or 60 to 69 (male), after which frequencies decrease again.

Daegu appears to have a higher number of "younger" female confirmed cases (ages below 59) but a higher number of "older" male confirmed cases (60 or older). All other provinces have so few cases, it's difficult to make any meaningful comparisons.

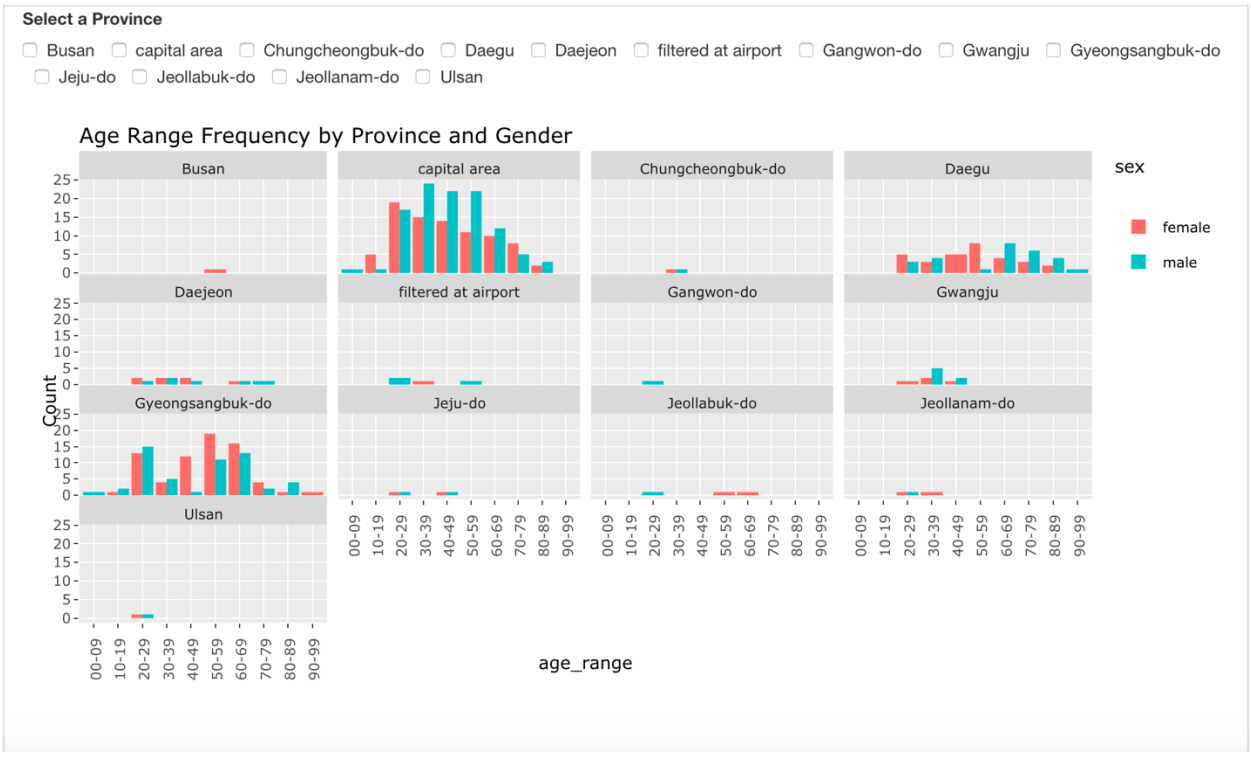

FIGURE 6. Grouped vertical bar chart. Examines frequency of confirmed cases by age (as 10-year bins) and gender frequencies in each province.

Daily Confirmed Cases by Province Line Chart. To examine the running total of confirmed cases over time in each province, a line chart was created using Plotly (see Figure 7). A line chart is an appropriate selection when plotting changes over time. Here, dates range from January 1st, 2020 to March 8th, 2020 and are presented along the x-axis. The y-axis represents the number of confirmed cases. Each province is plotted over time, identified on the visualization with a different color and marker type. The color-coded legend is provided for clarity. As the

chart was created in Plotly, users can trace changes over time and are provided tooltip information regarding count and province for each data point. Users can also double-click on the province legend to isolate one location. In addition, a province filter was created and displayed on a drop-down menu. Here, users can select one, multiple, or all provinces for comparison. It should be noted that the dataset included 21 "NA" occurrences in the province category. These null provinces yielded abnormally high numbers of confirmed cases, which dramatically altered the y-axis of this chart and make it too difficult to detect differences from named provinces. As a result, all "NA" provinces were eliminated for this visualization.

Upon inspection of the visualization, we see that the first observed cases occurred at the airport on January 20th and 24th, 2020. On January 26th, the capital area had its first confirmed case. Jeollabuk-do was the second province to have a confirmed case, observed on January 31st, 2020. Over all provinces, cases remained low until February 18th, 2020. At this time, Daegu had the first spike in confirmed cases, which dropped back down to 1 case shortly afterward (February 22, 2020) and remained low for the rest of the comparison.

Gyeongsangbulk-do and the capital areas both had multiple spikes during the observed time frame. Gyeongsangbulk-do observed a sharp increase in confirmed cases on February 22nd, 2020, a decrease in cases on February 23rd and 24th, a small increase on February 25th, the steepest observed increase on February 26th, and a dramatic decrease on February 27th, 2020. Confirmed cases remained relatively low thereafter. The capital area spikes never reached the maximum height observed in Gyeongsangbulk-do's; although, it had 5 increases in confirmed cases, which began after Gyeongsangbulk-do's first spike and continued after its last decline.

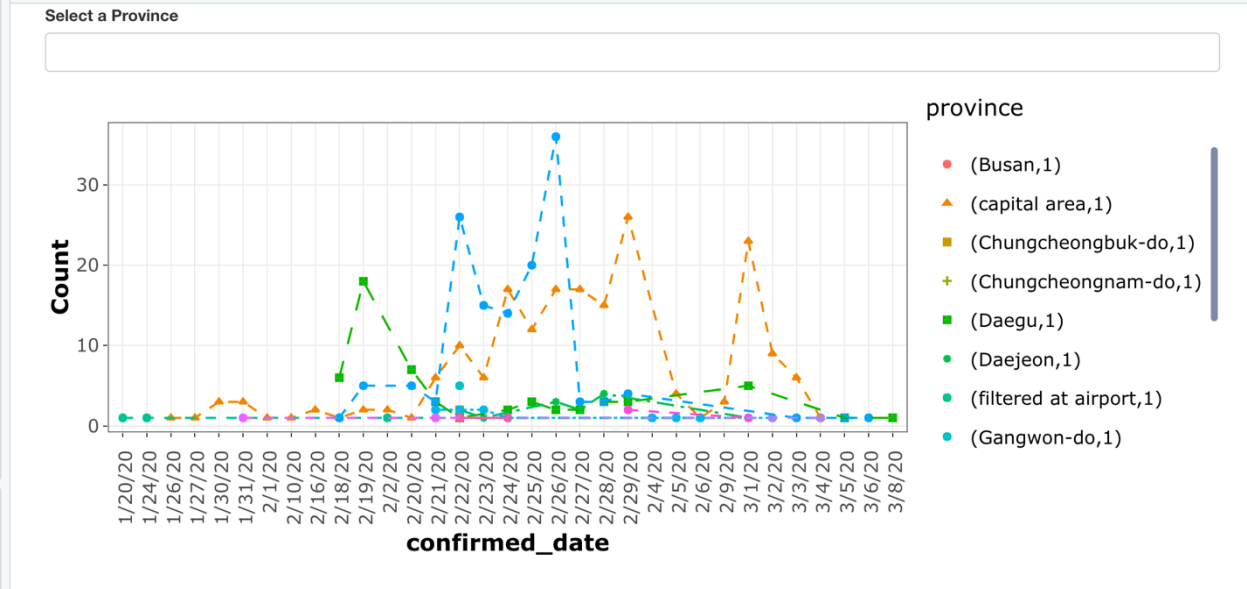

FIGURE 7. Line chart showing the number confirmed cases by day in each province in South Korea

Daily Infected, Recovered, and Deceased Case Numbers Line Chart. The chart above examined the daily number of confirmed cases within each province of South Korea (see Figure 8). To gather information regarding incubation periods, infection rates, recovery rates, and mortality rates, a line chart comparing number of cases in each "group" (confirmed cases, released cases, deceased cases) was compared across time with Plotly (see Figure 8). This data represents cases observed in all provinces, including the 21 "NA" provinces previously omitted from analysis.

Here, dates range from January 1st, 2020 to March 10th, 2020 and are presented along the x-axis. The y-axis represents the number of cases. Each group is plotted over time, identified on the visualization with a different color and marker type. The color-coded legend is provided for clarity. The marker type is size-dependent, such that higher number of cases have a larger marker size. This may be unnecessary, as only confirmed case group-types showed significantly large

numbers, making them visibly different from the other group types without the added size feature.

As the chart was created in Plotly, users can trace changes over time and are provided tooltip information regarding count and group for each data point. Users can also double-click on the group legend to isolate one group-type. In addition, a group filter was created and displayed on a drop-down menu. Here, users can select one, multiple, or all group-types for comparison. Additionally, filter check boxes were created so users could isolate one or multiple dates for comparison. Future analyses may opt to include a sliding bar instead of a check box as it's functionality may be a little more intuitive for users.

Visual inspection of the line chart indicates that both released and deceased case numbers remained low for the duration of this analysis. Confirmed cases, on the other hand, fluctuated greatly. Due to the fact that confirmed cases stayed elevated between February 2nd, 2020 and March 9th, 2020, one might consider the possibility that the recovery period lasts longer than one month.

Ostensibly, the number of confirmed cases gradually increases from February 2nd to February 29th and then drops down completely. However, this is a flaw in the data, as we see the decline is dated February 4th, 2020. Despite my best attempts, I was unable to fix this axis error. The visualization should read a spike in confirmed cases on February 9th, a steep decline on February 10th, a gradual increase between February 20th to February 29th, and then a gradual decrease until March 10th.

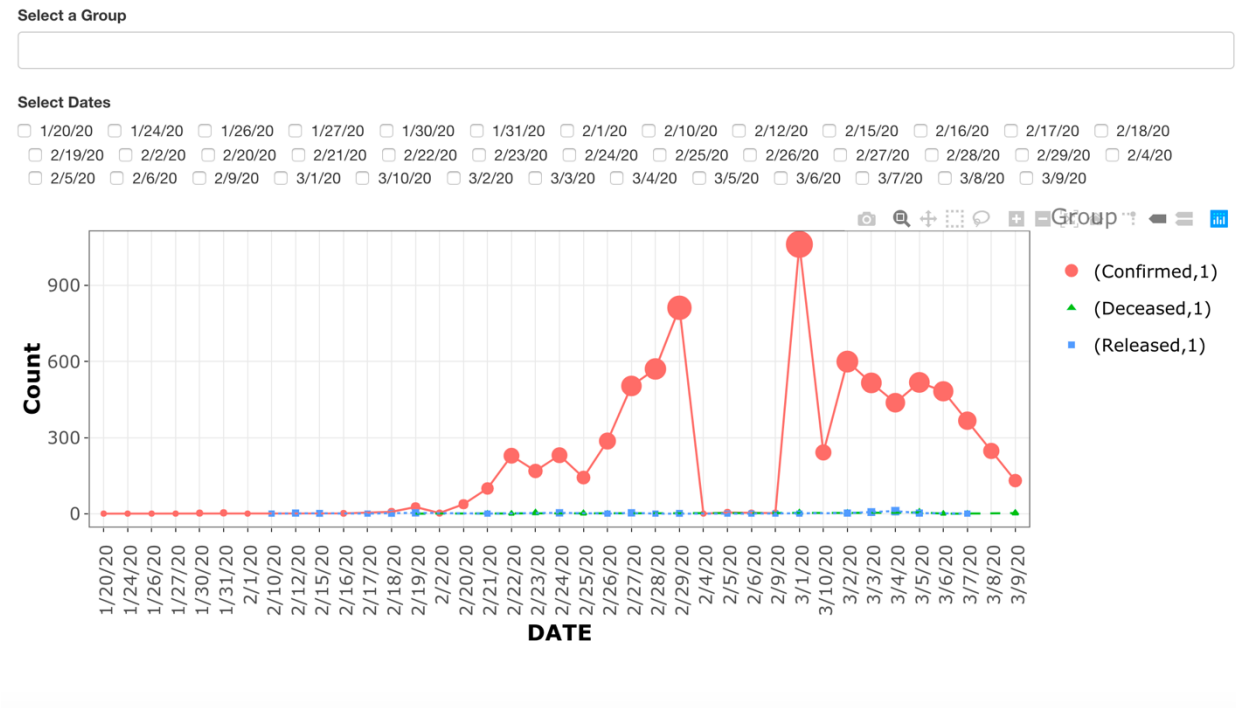

FIGURE 8. Line chart showing the number confirmed cases, deceased cases, and released cases by day in South Korea

Province by Infection Reason Heatmap. Thus far, analyses have examined patient demographics (age, gender, province) for confirmed COVID-19+ patients, as well as daily case numbers over time and by province. The question still remains how patients are spreading the disease. To address this question, a heatmap (or highlight text table) was created examining the frequency of patient reported infection reasons within each province (see Figure 9).

The heatmap creates a matrix of grids or cells. Here, columns represent levels of one categorical variable (e.g., province) whereas rows represent levels of another categorical variable (e.g., infection reason). The grid acts as coordinates between two discrete joint distributions. Each cell has an equal width and is filled along a continuous color scale, depending on the value of the relationship between two discrete variables. In a typical heatmap, the entire cell is filled by color. Thus, color strength is the only factor indicating the strength of a relationship (i.e., size

and shape are held constant). Due to the fact that color is proportional to the value of the relationship, viewers can quickly identify incidence patterns. Here, the frequency (or count) of infection reasons is tallied and labeled in each grid space, such that the viewer can see how many people in each region indicated the specified infection reason. The stronger the color of the cell (i.e., stronger in shade, color, or brightness), the stronger (or larger) the higher number of individuals reported that infection reason.

Here, the heatmap was created with Plotly, which uses a user-driven trace function, allowing users to interact with each cell block. Each cell provides a tooltip labeling the province, infection reason, and count/frequency. For clarity, a title was added to the heatmap and a color legend was provided.

Overall, we see that most of the chart is shaded in a light blue, indicating low frequencies for most infection reasons. Considering many provinces had low numbers of confirmed cases, this is expected. The highest reason reported was in the capital area for “contact with patient”, followed by “visit to Daegu” for patient living in the province Gyeongsangbulk-do. “Contact with patient” is the highest infection reason presented (76 cases), followed by “visit to Daegu” (50 cases). Taken together, these results are interesting given the information obtained from Figure 7, which suggested Daegu as the first province to have a spike in the number of confirmed cases. The provinces of Daegu, Gwangju, Gyeongsangbulk-do, and the capital area had much more variety in infection reasons.

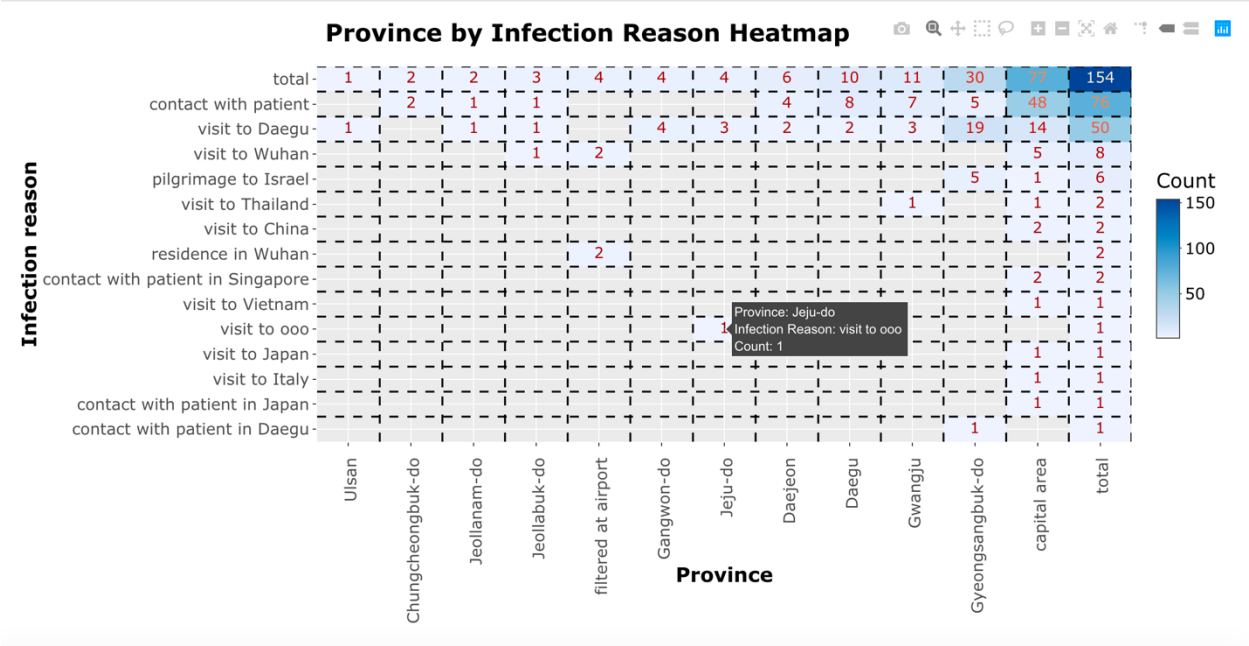

FIGURE 9. Province by infection reason heatmap. Each value in a cell indicates the number of patients in the specified province (column) who reported the corresponding infection reason (row).

Province by Infection Reason Grouped Vertical Bar Chart. To address some of the issues with readability with the heatmap, a Plotly grouped vertical bar chart was also created to compare province by infection reason (see Figure 10). Here, province was represented on the x-axis and count (or, frequency) on the y-axis. Each infection reason was grouped by color and was represented as the color-fill of the vertical bar within each province. Additionally, a province filter check box was provided to give users more control. In this way, users could double-click infection reason legend markers to isolate specific patient reported infection reasons and select one or many provinces for comparison.

Overall, conclusions made while viewing the heatmap were consistent with that viewed on the grouped vertical bar chart. It is suspected that the grouped vertical bar chart may be more intuitive for some users.

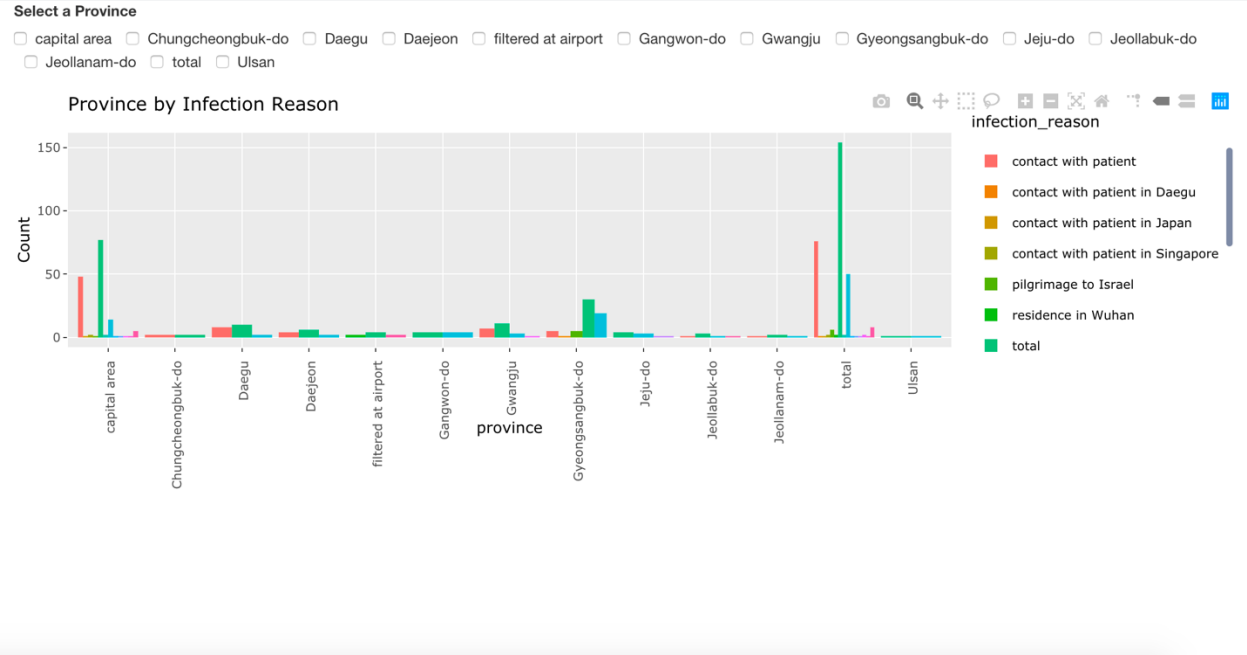

FIGURE 10. Province by Infection Reason Grouped Vertical Bar Chart.

Description of Project 3: FIFA 2019

This project served as a sports analysis for the Federation Internationale de Football Association (FIFA)'s 2019 player roster. The data set includes 89 variables, including player names, age, nationality, overall, potential, value, wage, international reputation, and multiple skill-specific sport factors. By identifying key factors and interactions between these variables, we might better understand the international appeal, value of, and skill-set of players and teams within FIFA.

Data Set

The data set, FIFA 19 Complete Player Dataset, was obtained from Kaggle (https://www.kaggle.com/karangadiya/fifa19). The data set included a total of 89 variables. Rather than using them all, only the following were selected: name, age, overall, potential, value, wage, nationality, international reputation, RF, ST , LW, GK, RCM, LF, RS, RCB, LCM, CB,

LDM, CAM, CDM, LS, LCB, RM, LAM, LM, LB, RDM, RW, CM, RB, RAM, CF, RWB, LWB, NA, GK, Crossing, Finishing, Heading, Accuracy, ShortPassing, Volleys, Dribbling, Curve, FKAccuracy, LongPassing, BallControl, Acceleration, SprintSpeed, Agility, Reactions, Balance, ShotPower, Jumping, Stamina, Strength, LongShots, Aggression, Interceptions, Positioning, Vision, Penalties, Composure, Marking, StandingTackle, SlidingTackle, GKDiving, GKHandling, GKKicking, GKPositioning, and GKReflexes.

Positions were grouped into the following classes: defense, midfielder, goalkeeper, and forward. The defense class included the following variables: CB, RB, LB, LWB, RWB, LCB, and RCB. The midfielder class included the variables: CM, CDM, CAM, LM, RM, LAM, RAM, LCM, RCM, LDM, and RDM. The forward class included the variables: RF, ST, LF, RS, LS, RW, and CF. The goalkeeper class included the variable GK. All position variables were character datatypes representing a rating on a scale of 100.

The following variables all represent specific soccer skills, values given on a rating on scale of 100 (numerical datatype): Crossing, Finishing, Heading, Accuracy, ShortPassing, Volleys, Dribbling, Curve, FKAccuracy, LongPassing, BallControl, Acceleration, SprintSpeed, Agility, Reactions, Balance, ShotPower, Jumping, Stamina, Strength, LongShots, Aggression, Interceptions, Positioning, Vision, Penalties, Composure, Marking, StandingTackle, SlidingTackle, GKDiving, GKHandling, GKKicking, GKPositioning, and GKReflexes.

Using R, descriptive statistics were conducted on all variables in order to obtain their frequency, mean, median, mode, minimum, maximum, ranges, standard deviation, variance, skew, and kurtosis values. Exploratory data analysis utilized scatterplots, histograms, and boxplots for individual variables to better understand their central tendencies, variability, and to detect outliers. Exploratory data analyses were conducted in R's ggplot, Plotly, and highchart.

Though not for the purposes of the required visualizations of this assignment, the following were created for exploratory purposes: overall rate histogram (highchart and plotly), overall rate histogram comparison (normal, Sturges, and Scott histograms in plotly), normal Q-Q plot for overall rate, potential rate (highchart and plotly), normal Q-Q plot for potential rate, age histogram (highchart and plotly), age histogram comparison (normal, Sturges, and Scott histograms in plotly), overall by age boxplot (highchart), potential by age boxplot (highchart), wage by overall boxplot (highchart), wage by potential boxplot (highchart), wage by potential vertical line chart (highchart), value by overall boxplot (highchart), value by potential boxplot (highchart), value by potential vertical line chart (highchart), overall rate by top player polar/circular bubble chart, number of players choropleth (ggplot2), correlation heatmap (plotly), and p-value correlation heatmap (plotly). Code and screenshots for these accessory visualizations are provided in the .zip file.

Visualizations and Updates

Top Player by Overall Rate and Nationality Scatterplot. In order to visualize the top players by their overall rating, a scatterplot was created (see Figure 11). Thirty players with the highest Overall ratings were selected. Their names were represented on the x-axis and overall rating was represented on the y-axis. The marker for each player's overall score was scaled by size, such that players with higher overall ratings had larger marker sizes. Each player was further categorized by nationality, indicated by a color-coded marker. The visualization was originally designed as a polar scatterplot, so that lower overall values would cluster closer to the center of circle (see Figure 12), but the interactive components required for this assignment did not translate well to this format. Therefore, a simple scatterplot was utilized.

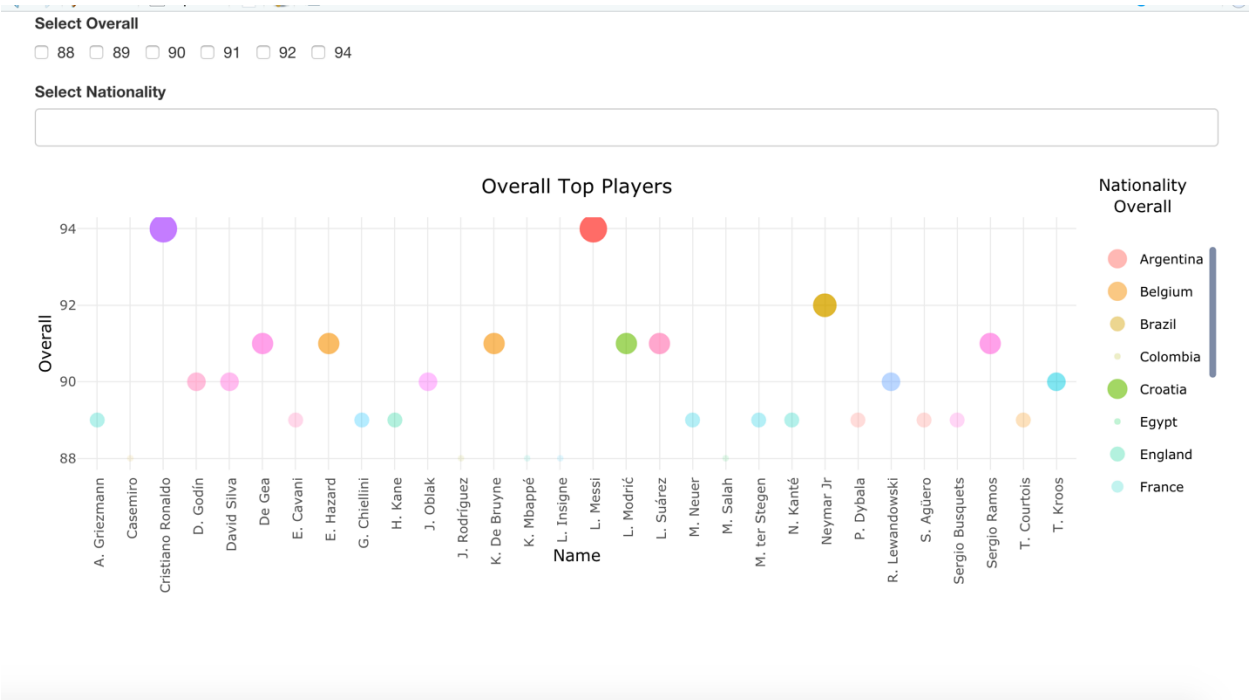

FIGURE 11. Top players by Overall and Nationality scatterplot. Overall filter checkboxes, nationality filter drop-down menu and Plotly functionality are included.

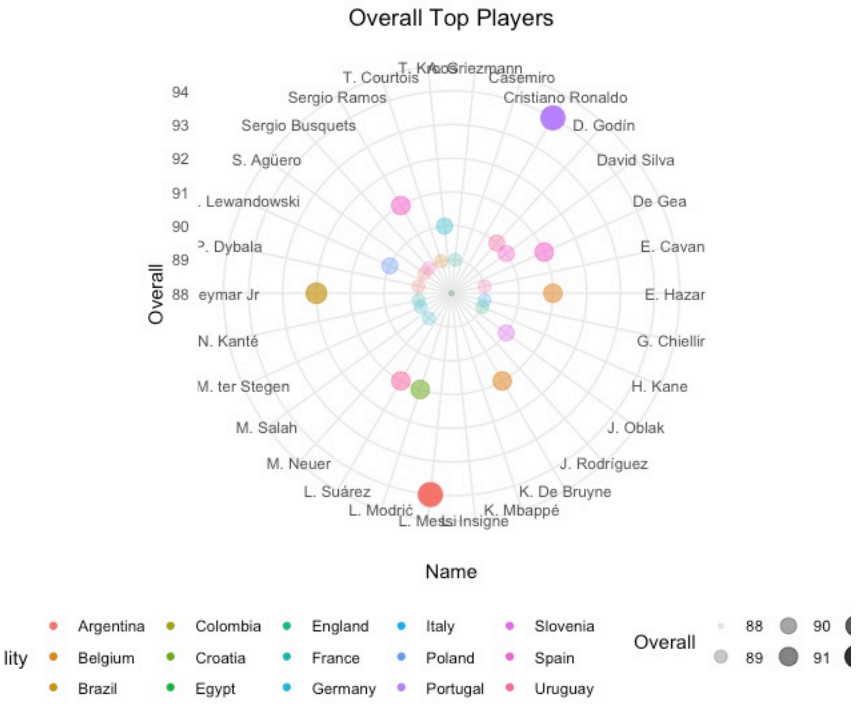

FIGURE 12. Top players by Overall and Nationality polar scatterplot. The original design for this visualization was a polar scatterplot. However, user-interactivity was too difficult to include in a polar design.

By inspecting this visualization, we see two players with the highest overall scores: Cristiano Ronaldo (Rating: 94, Nationality: Portugal) and L. Messi (Rating: 94; Nationality: Argentina). One player has an overall rating of 92 (Neymar, Jr; Nationality: Brazil), 6 players have an overall rating of 91, 4 players have an overall rating of 90, 11 players have an overall rating of 89, and 5 players have an overall rating of 88. Spain has the most players in the top by overall, followed by Belgium. Although, it should be noted that with so many countries included, the color scale makes certain markets look too similar to one another (e.g., it's very difficult to tell the difference between the colors of Spain and Uruguay).

Top Player by Potential Rate and Nationality Scatterplot. As we know, potential ratings do not always equal overall ratings. In fact, a potential rating could be used to predict future performances. As a result, a similar design to Figure 11 was used to create another vertical scatterplot visualizing top players by their potential rating and nationality (see Figure 13). Data point markers were grouped by size and nationality, such that nationality was color-coded and marker size was proportional to the size of the potential rating. As before, the visualization was created in Plotly, such that users have the option to select a specific nationality. A potential rate filer checkbox was included, as was a nationality filter drop-down menu. For variety, however, the axes were switched, such that the names of each player were represented on the y-axis and their potential ratings on the x-axis.

We see that there is one player with a potential rating of 95 (K. Mabappe; Nationality: France), 4 players with potential ratings of 94, 4 players with potential ratings of 93, 9 players with potential ratings of 92, 12 players with potential ratings of 91, and 1 player with a potential rating of 90. The same issues with color and nationality present here. It is difficult to differentiate some countries.

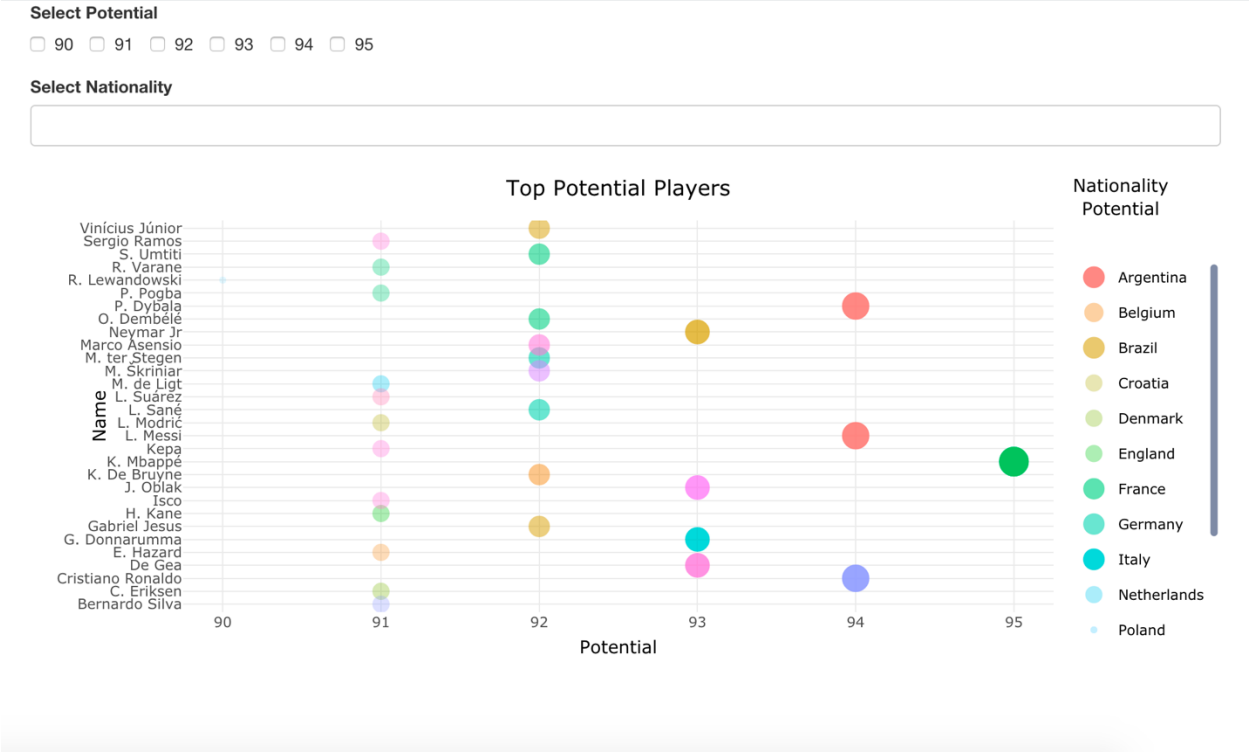

FIGURE 13. Top player by potential rate and nationality scatterplot.

Skillset Correlation Heatmap. It's known that each player in the FIFA has an overall rating as well as six scores for the key stats (Pace, Shooting, Passing, Dribbling, Defending, and Physical), which are combined with a player's international recognition to calculate their overall rating. As we've already examined overall rating, the next step would be to start looking at skillsets. As the data set includes a huge number of skillsets, the first step in analysis is to examine their correlations. Thus, a skillset correlation heatmap was created in Plotly (see Figure 14).

This heatmap used a continuous color scale, ranging from dark blue (indicating a strong positive correlation) to dark red (indicating a strong negative correlation). The use of Plotly's user-driven trace function is key here, as there are many variables for comparison. As mentioned earlier regarding critiques of other heatmaps, this one also suffers from an abundance of color which doesn't help the eye focus on any one spot in particular. However, the user can explore interested variables and/or seek out abnormally dark or light colors.

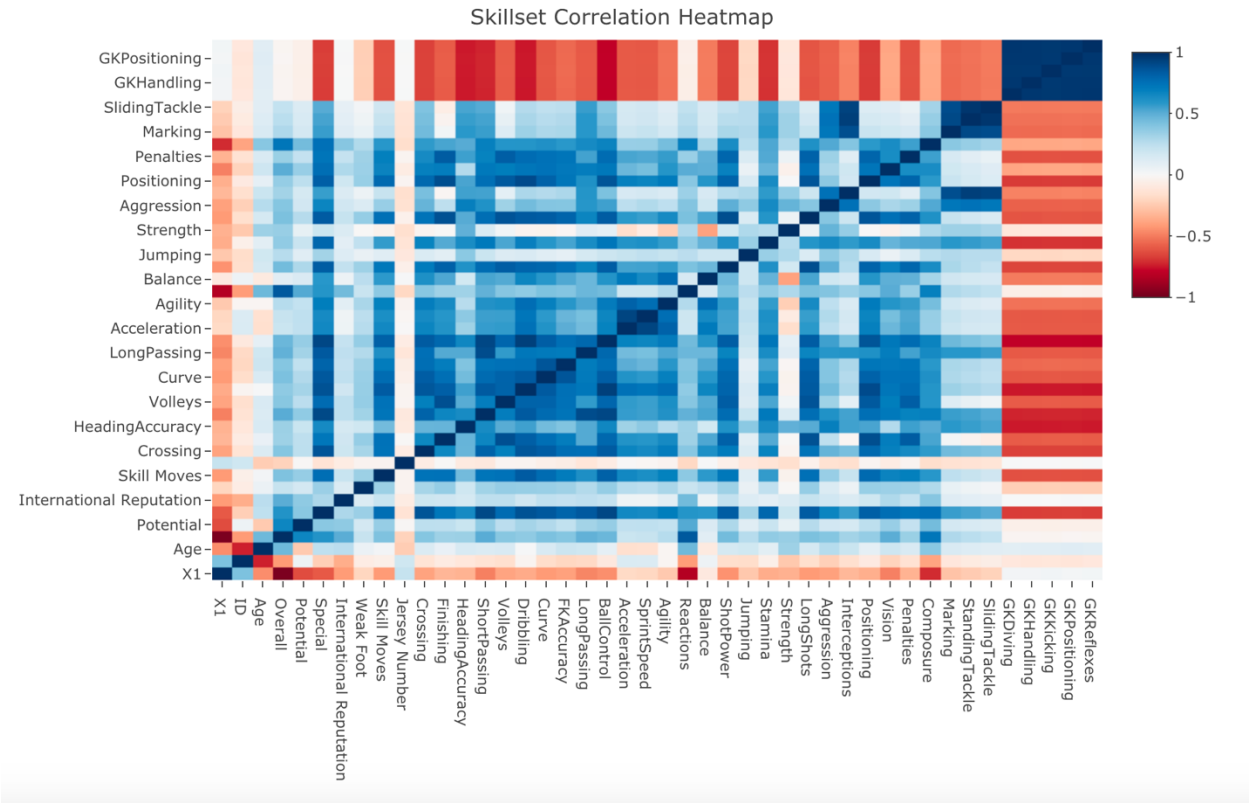

FIGURE 14. Skillset correlation heatmap.

Visual inspection of Figure 14 reveals a few interesting trends. Overall has a strong positive correlation with Reactions, Potential is not correlated any variables examined, International Reputation is not correlated with any variables examined, Sliding Tackle has a strong position correlation with Interceptions, Ball Control has a weak negative correlation with Goalie Handling, Sprint Speed has a strong positive correlation with Acceleration, and many more observations.

Skillset Correlation p-Values Heatmap. To visualize only significant correlations, the same heatmap was also visualized using a matrix of p-values (see Figure 15). This makes it easier to focus on relevant relationships. However, in retrospect, it would have been better to keep both correlation heatmaps on the same color-scale.

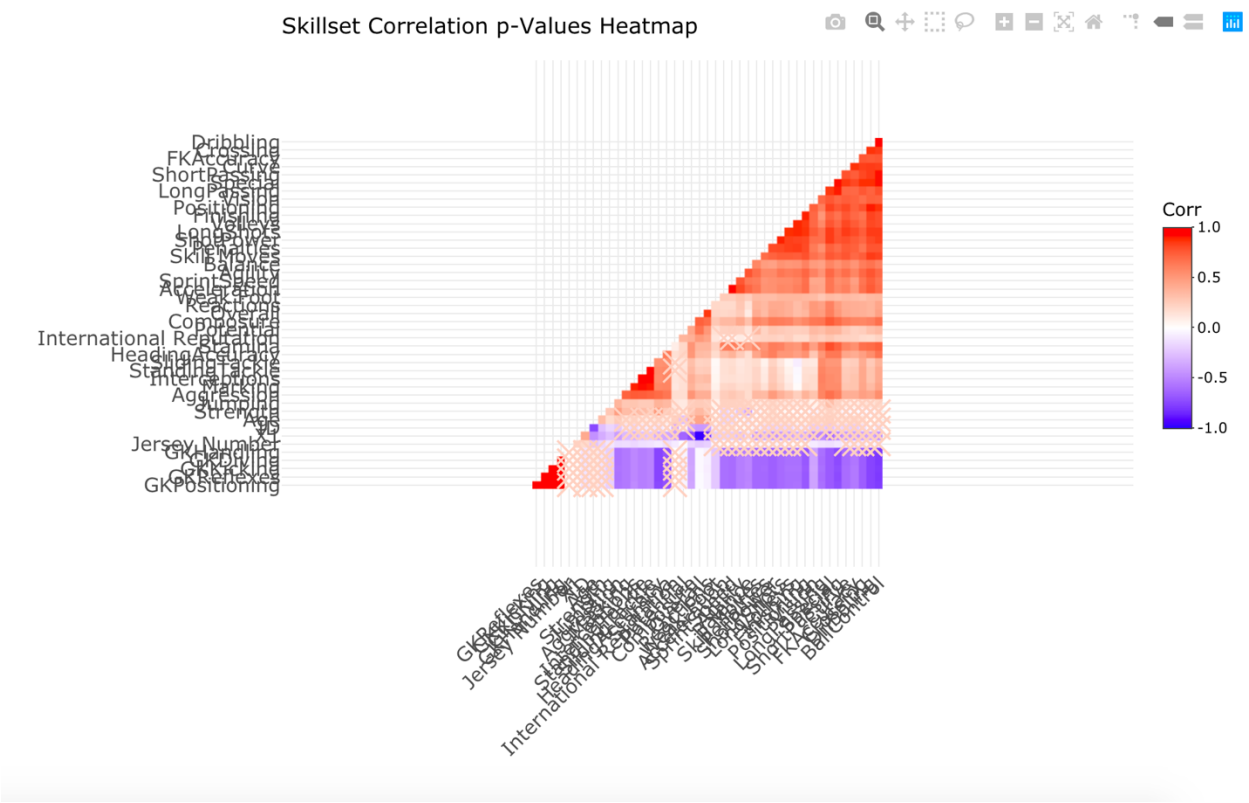

FIGURE 15. Skillset correlation p-value heatmap.

Position by International Reputation Scatterplot. Since we now know that international reputation is not correlated with any specific skillset, we may want to look at the relationship between it and player position. To do so, a position by international relationship scatterplot was created in Plotly (see Figure 16). Each player position was represented on the x-axis while international reputation was represented on the y-axis. Due to the large number of positions, each datapoint is color coded by position class (defender, midfielder, goalie, forward). Plotly allows the user to select one class for comparison. An attempt was made to also include a position filter drop-down menu (see Figure 17), but this greatly affected the scatter jitter and the x-axis theme, making it very difficult to read. Overall, we see that the class positions with the highest international reputation (4) are forwards and goalies. All other reputations (3, 2, and 1) have a more even distribution of positions and position classes.

FIGURE 16. Position by international reputation scatterplot.

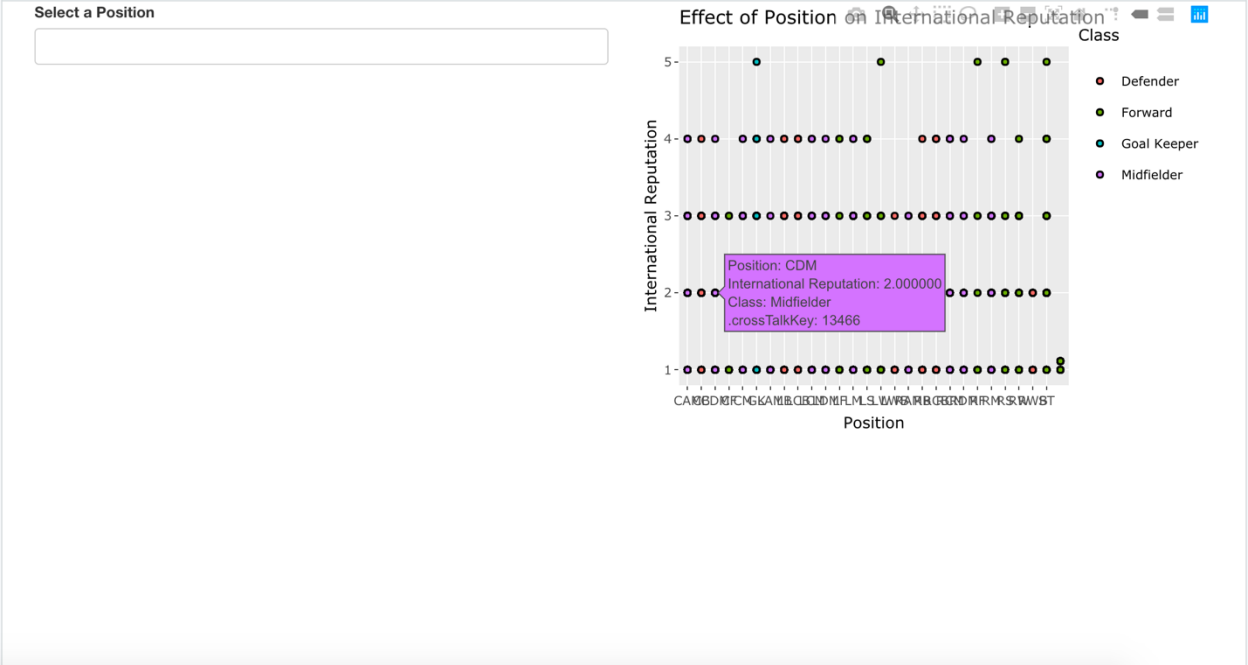

FIGURE 17. Position by international reputation scatterplot with position filter drop-down menu.

Conclusions

Interactive visualizations allow users an opportunity to explore and engage with data in ways that static graphs and charts are unable. This project utilized R's Plotly and HighChart packages to provide user-driven interactions for three data sets. It was observed that the inclusion of interactive components required more specific chart formatting and created unexpected trouble with axes and special plot-types. Redundant features such as Highlights and Filters used through sliding scales, checkboxes, or drop-down menus should be considered against the functionality built into packages such as Plotly.

References

Kim, J. (n.d.). Data Science for COVID-19. Retrieved December 01, 2020, from https://www.kaggle.com/kimjihoo/coronavirusdataset?select=Case.csv

Gadiya, K. (n.d). FIFA 19 Complete Player Dataset. Retrieved December 01, 2020, from https://www.kaggle.com/karangadiya/fifa19

Millais, P., Jones, S. L., & Kelly, R. (2018, April). Exploring data in virtual reality: Comparisons with 2d data visualizations. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1-6).

Skidmore, Z. L., Wagner, A. H., Lesurf, R., Campbell, K. M., Kunisaki, J., Griffith, O. L., & Griffith, M. (2016). GenVisR: genomic visualizations in R. *Bioinformatics*, *32*(19), 3012-3014.

Seshapanpu, J. (2018). Student Performance in Exams. Retrieved November 30, 2020, from https://www.kaggle.com/spscientist/students-performance-in-exams/discussion/160544