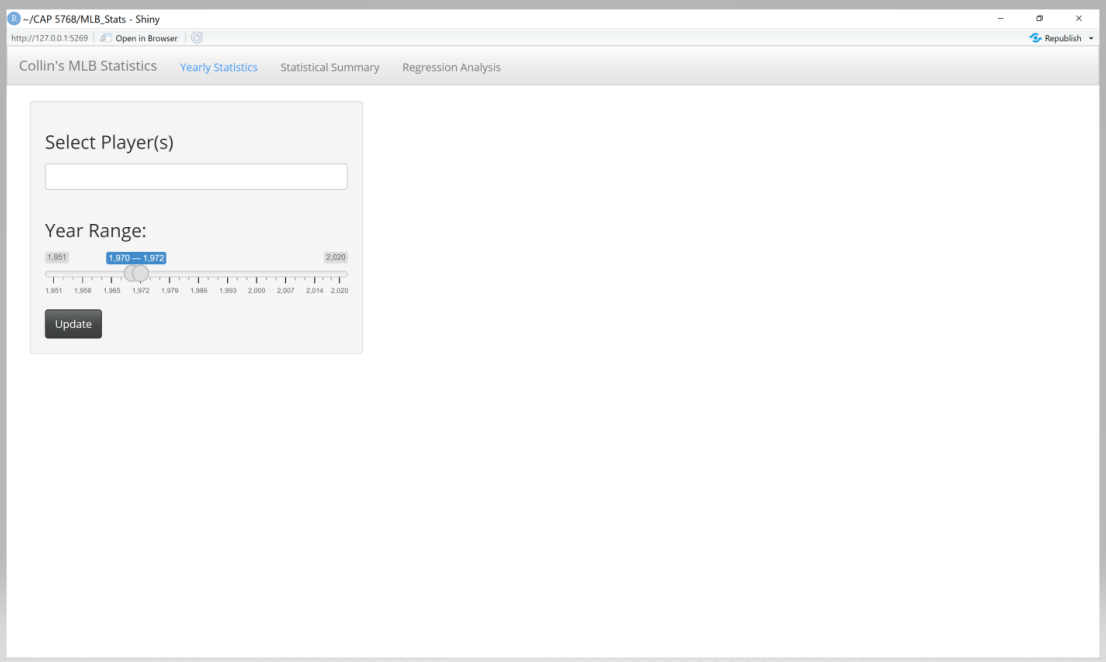
1. Introduction
   MLB_Stats is a R Shiny application that allows users with limited baseball knowledge to view statistics and conduct analysis with multiple common baseball statistics. This document serves the purpose of giving users a quick preview of using the application and explains the meaning of the statistics to the user. The MLB_Stats is a public site that can be viewed at https://github.com/cbakken2021/MLB_Stats. Each of the three modules included in the interface will be briefly discussed throughout this walkthrough.

2. How to Start
   In order to properly run MLB_Stats locally, users will need the following at their disposal.
   https://github.com/cbakken2021/MLB_Stats
   Requirement:
   a. R (version 4.0.2 or later)
   b. Shiny (version 1.2.0 or later)

   In order to install the Shiny package in R:
   a. Open R
   b. Run install.packages("shiny")

   Now, to run the MLB_Stats locally:
   a. Open R
   b. Run MLB_Stats by using the following commands in R:
      a. library("shiny")
      b. shiny::runGitHub("MLB_Stats", "cbakken2021", ref = "main")
   c. The image below will be the cover of the application that should appear when the code is executed properly.


~/CAP 5768/MLB_Stats - Shiny
http://127.0.0.1:5269
Open in Browser
Republish
Collin's MLB Statistics
Yearly Statistics
Statistical Summary
Regression Analysis
Select Player(s)
Year Range:
1,951
1,970 — 1,972
2,020
1,951 1,958 1,965 1,972 1,979 1,986 1,993 2,000 2,007 2,014 2,020
Update

Detailed instructions on how to run MLB_Stats:

(All statistical abbreviations are at the bottom of the document for the convenience of those who are unfamiliar with baseball terminology)

Module 1: Yearly Statistics

Module 1 can be used to view statistics of MLB players from 1951-2020. The user can either select a player from the dropdown menu or search for a specific player by typing into the menu located under “Select Player(s)”. The user can select multiple players in order to compare players against one another.

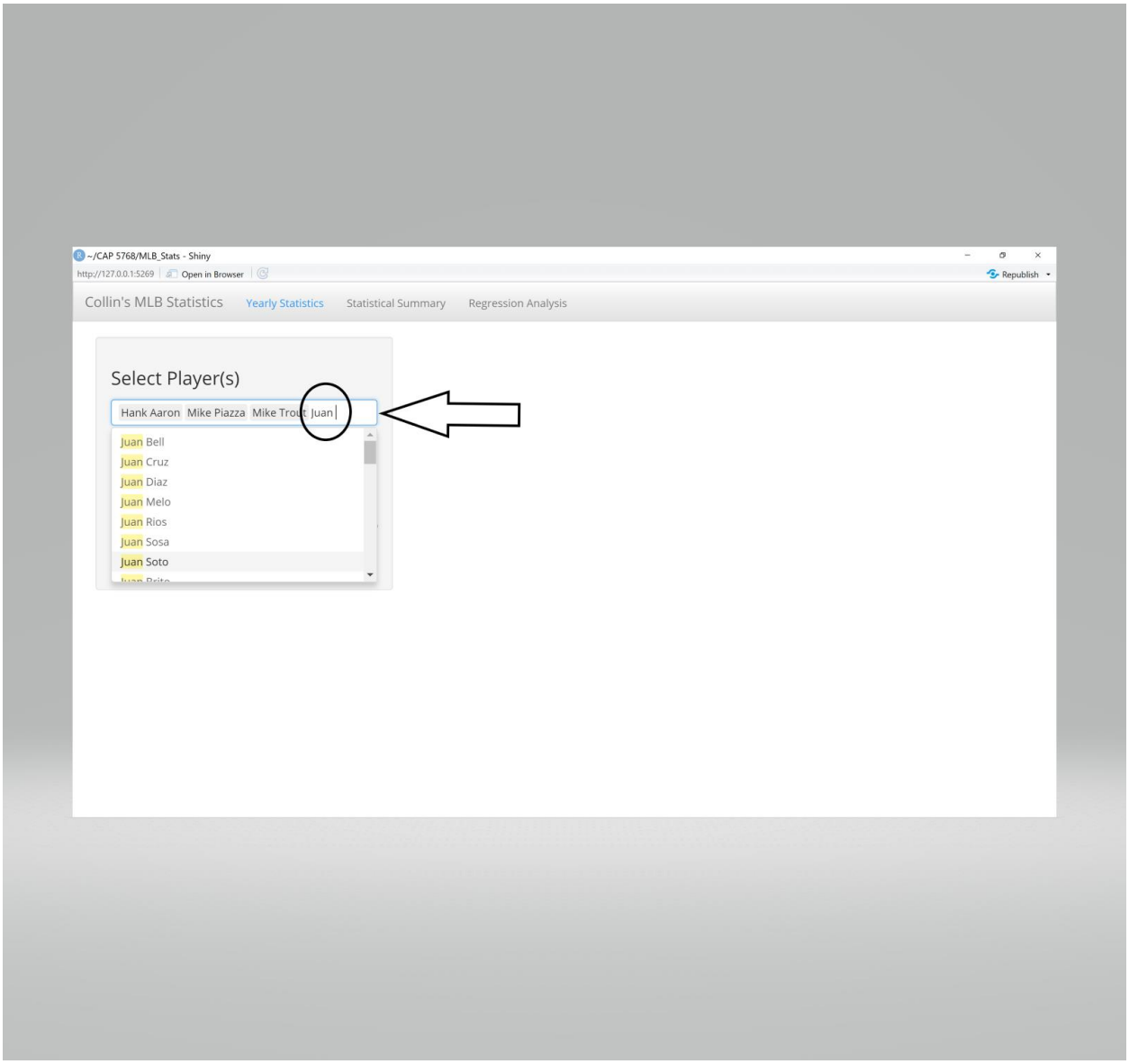


Figure 1: The above shows a user selecting multiple players (Hank Aaron, Mike Piazza, Mike Trout) and using the search interaction to search for Juan Soto.

The user is also asked to insert a year range in which to view these players stats. The user can use the entire year range (1951-2020) to see the full careers of players, or slim it down to as little as one year to view a player's stats throughout any specific year. Once all choices have been made, the user must hit "Update" and then the table will show all statistics desired in a table format as shown in figure 2.

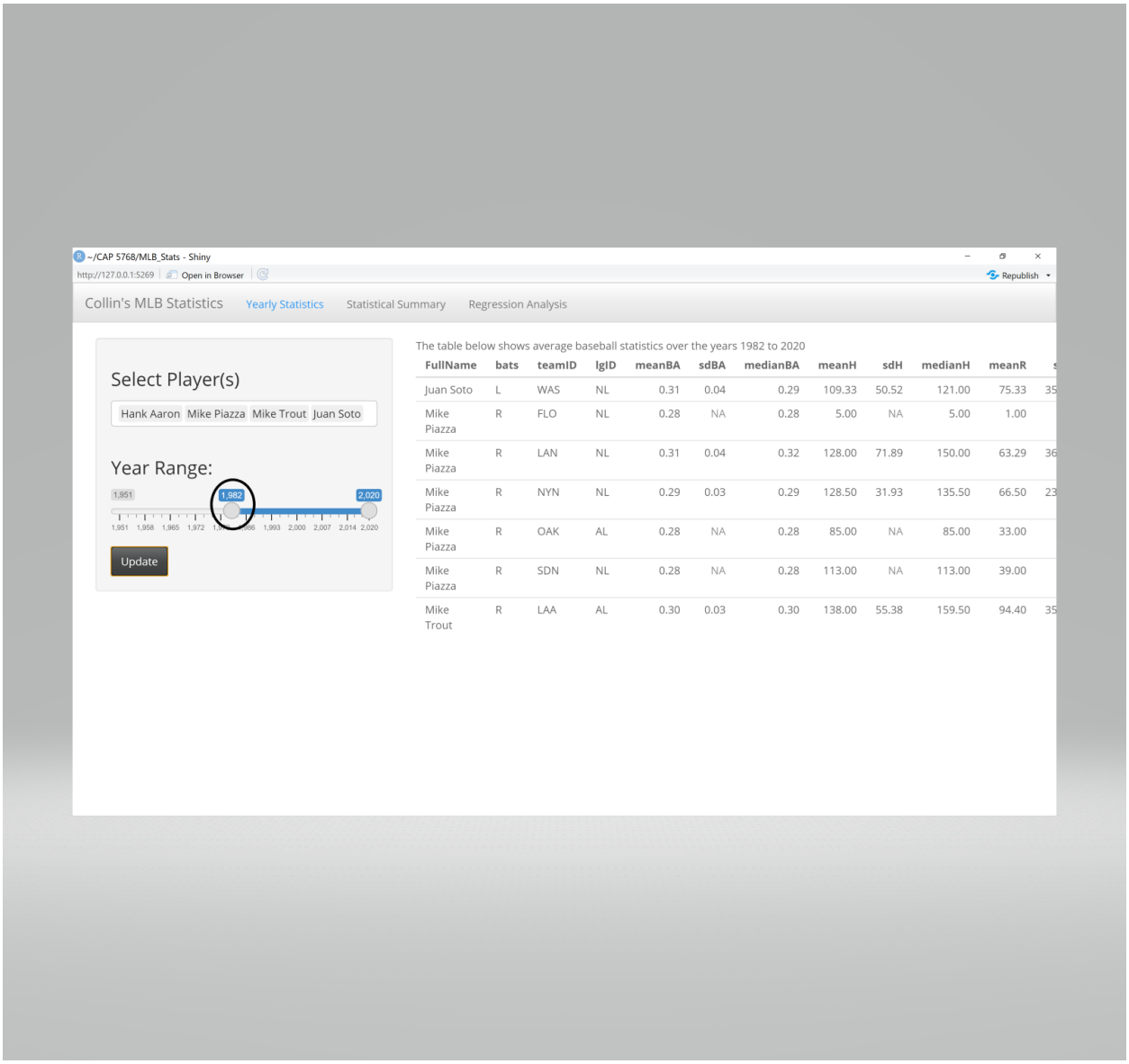


Figure 2: The user uses the scroll bar to select all statistics accumulated between 1982-2020. See that although Hank Aaron has been selected, no statistics are displayed, as he retired in 1976. (Note: If players selected do not have any statistics in the year range provided, the table will not display any data for the selected players.)

Module 2: Statistical Summary

Module 2 allows user to pick any teams they desire from a dropdown menu by the abbreviations that each teams use (for example, ATL = Atlanta Braves). Users can select as many teams as they wish, with a minimum of one team being required for displays to be created. Similar to module 1, the user also has access to a slider bar to use the year range they desire.

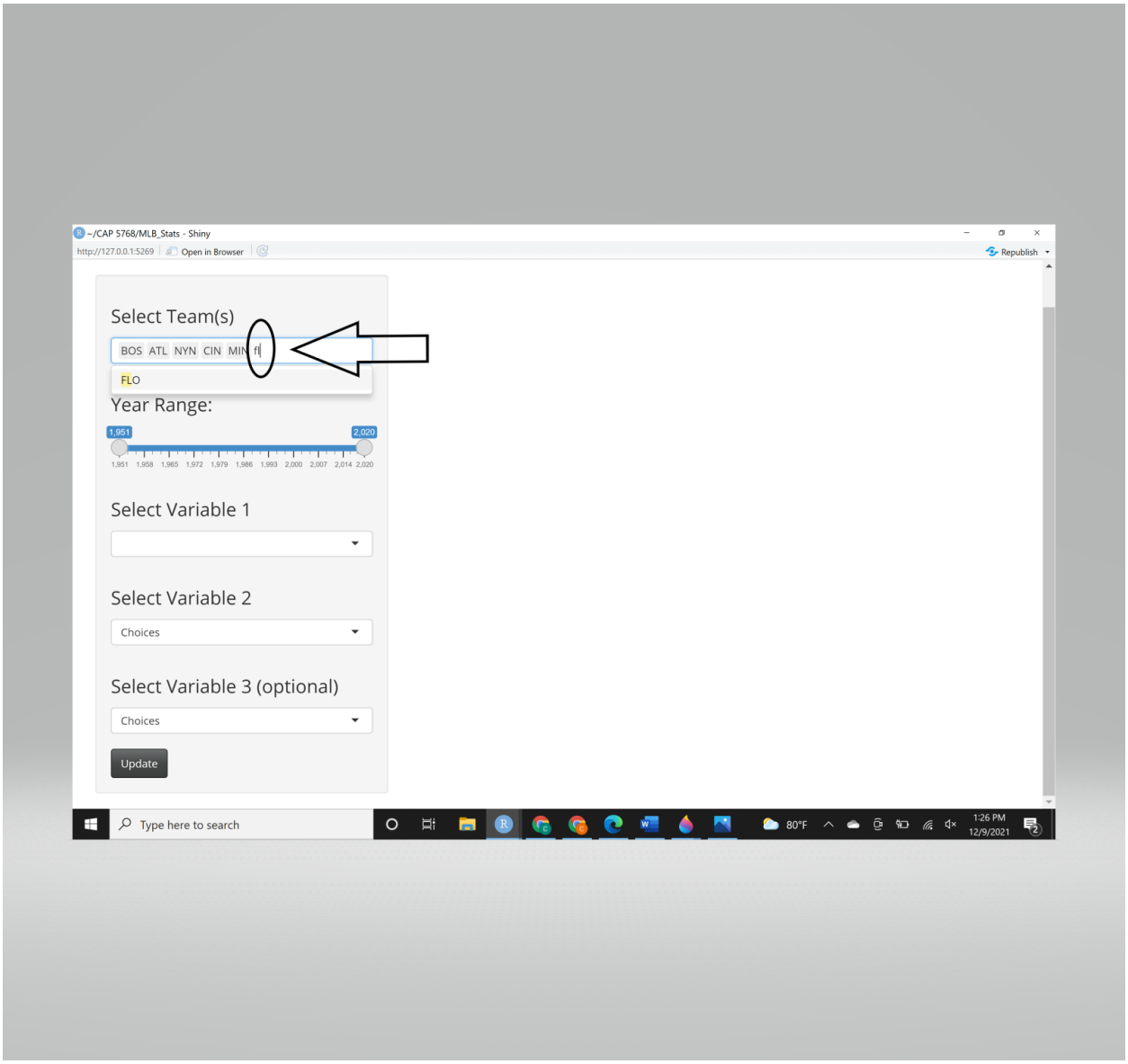


Figure 3: In the second module, our user has selected BOS (Red Sox), ATL (Braves), NYN (Mets), CIN (Reds), MIN (Twins), and is searching for FLO (Marlins). The user is using the maximum year range to see all displays provided in the dataset.

Once the user has completed selecting their teams and the year range, the user then must pick at least two statistics to compare against each other. A third choice is optional. The choices made by the user will determine the type of plot created based on if the user wants categorical or continuous variables. The statistics are shown in a dropdown menu under “Select Variable” and the user can choose from these options or search for a specific statistic. (Note: if “Select Variable 1” and “Select Variable 2” are left as the dummy variable “Choice”, no output will be displayed.)

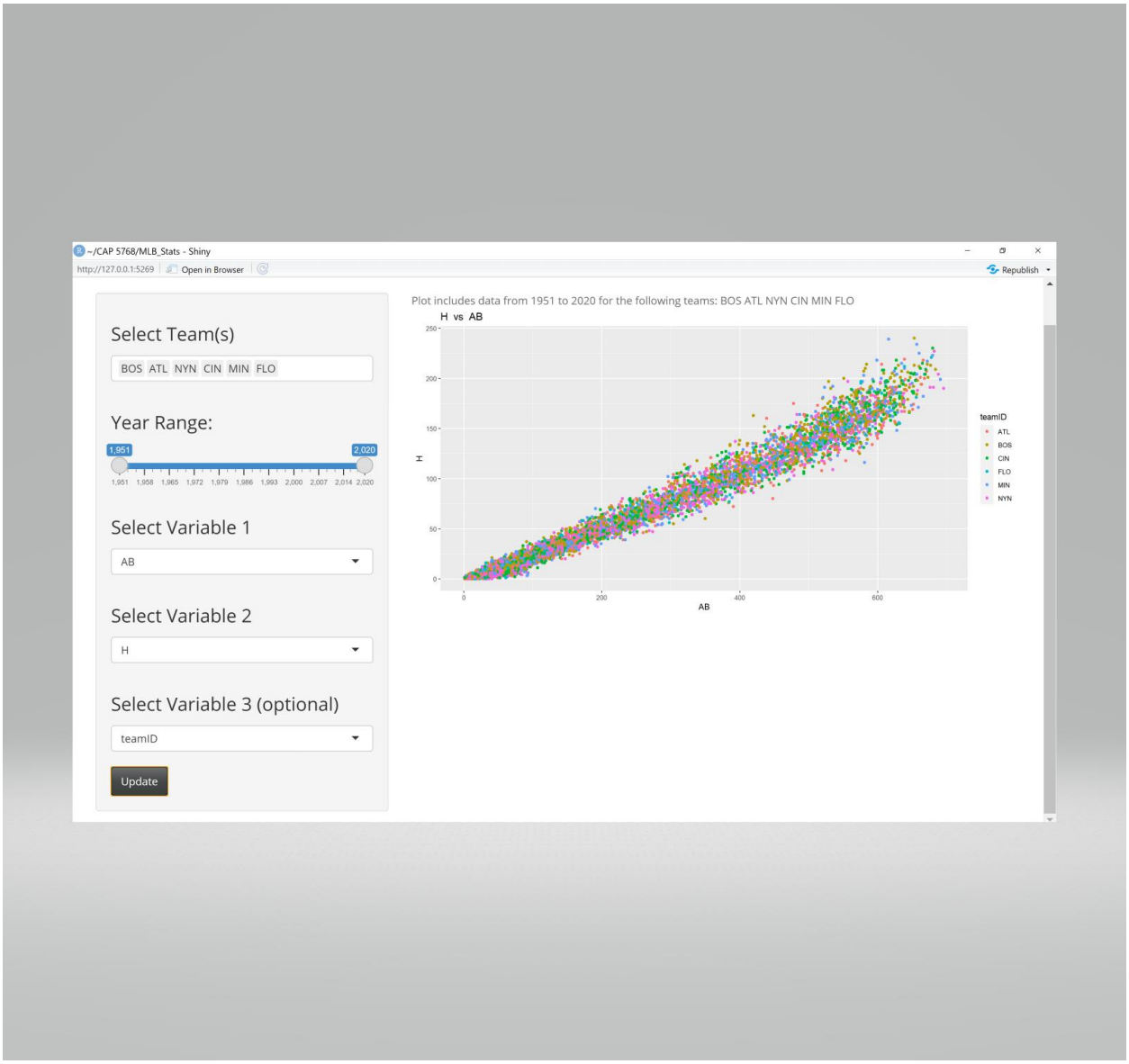


Figure 4: The user has selected two continuous variables (AB and H) and one categorical variable, teamID. The output is a scatterplot which shows Hits vs. At Bats with respect to different teams.

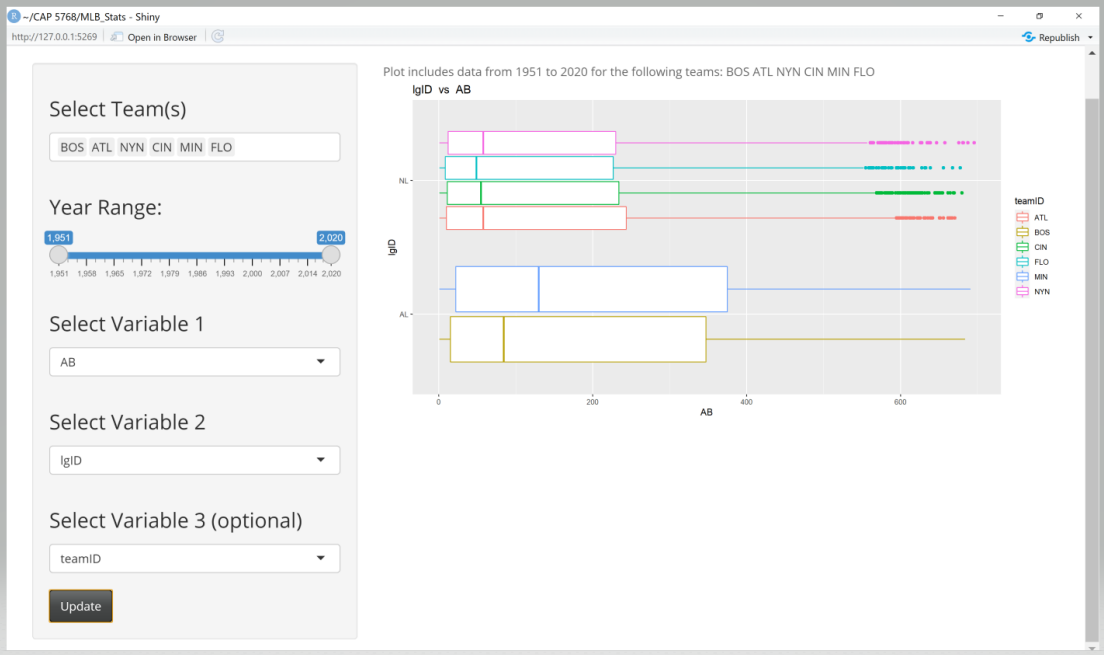


Figure 5: If a user elects two categorical variables (such as lgID and teamID) and a continuous variable (in this case, AB), then a boxplot will be displayed to see the correlation.

Module 3: Regression Analysis

Module 3 allows user to pick any statistic they desire to see how another variable (or a combination of variables) effects the response variable. If a user does not insert a variable into both the response and predictor(s), no display will be output. The module works best when two (or more) variables are used that have a meaningful correlation (for example, predicting AB based on teamID is not very practical). Once the user has filled in both variables and hits update, a display chart will show which gives the statistics of the analysis (given that there is a practical correlation) and a plot will show to visualize the linear regression model.

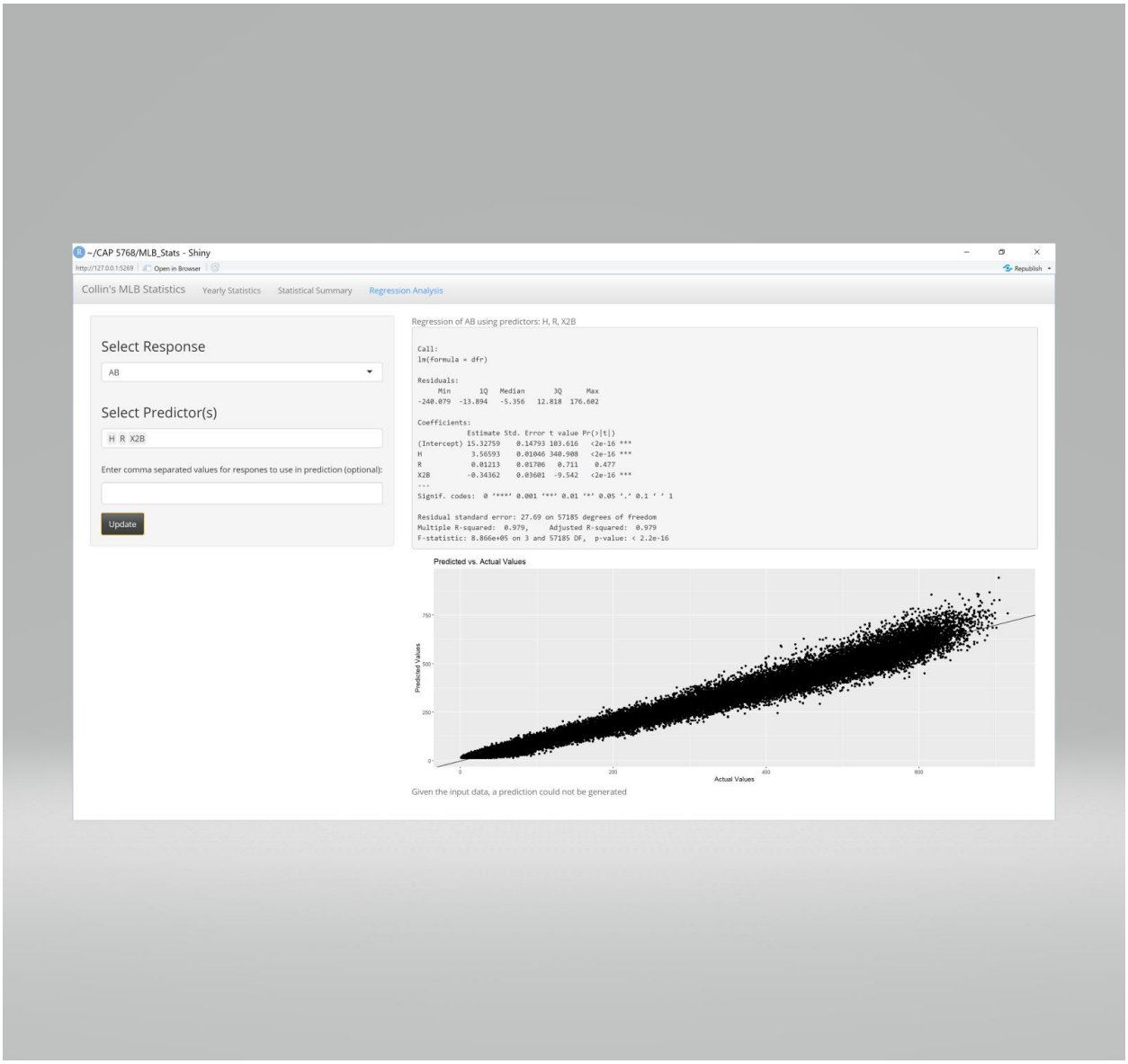


Figure 6: In this example, the user is attempting to predict at-bats based on hits, runs, and doubles. We can see there is a very strong correlation with a coefficient of 0.979. The estimates for each individual predictor are shown as well, with their p-values.

A user can also insert guesses per for each predictor variable in order to predict the outcome of the response variable. An example is shown below.

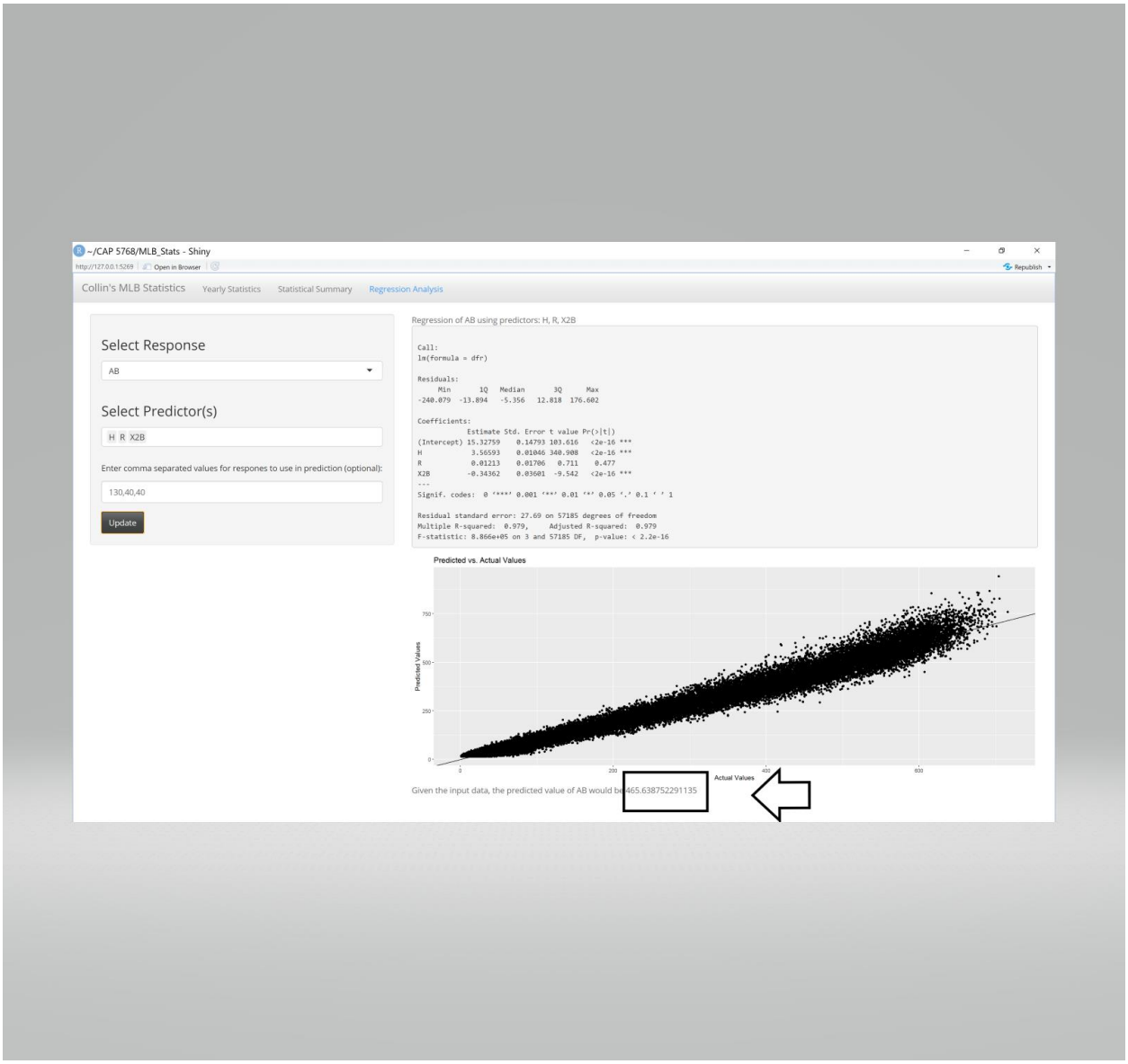


Figure 7: The user is predicting at-bats from hits, runs, and doubles. The user predicted if a player had 130 hits, 40 runs, and 40 doubles, the player had roughly 466 at-bats.

Abbreviated Statistic	Statistic	Brief Description
Weight	Weight	Weight of a player
Height	Height	Height of a player
Bats	Bats	The hitter hits right-handed (R), left-handed (L), or both (B)
yearID	yearID	The year that is being discussed
teamID	teamID	The team a player played for
lgID	lgID	National League or the American League
AB	At Bats	Number of at-bats
R	Runs Scored	Number of times a player scored
H	Hits	Number of hits a player had
X2B	Doubles	Number of doubles a player had
X3B	Triples	Number of triples a player had
HR	Homeruns	Number of homeruns a player had
RBI	Runs Batted In	Number of runs batted in a player had
SB	Stolen Bases	Number of stolen bases a player had
BB	Walks	Number of walks a player had
SO	Strikeouts	Number of strikeouts a player had
BA	Batting Average	The number of hits a player had divided by the number of at-bats the player had