DSCI 531 - LAB 1 - REPORT ON VISUALIZATION SURVEY

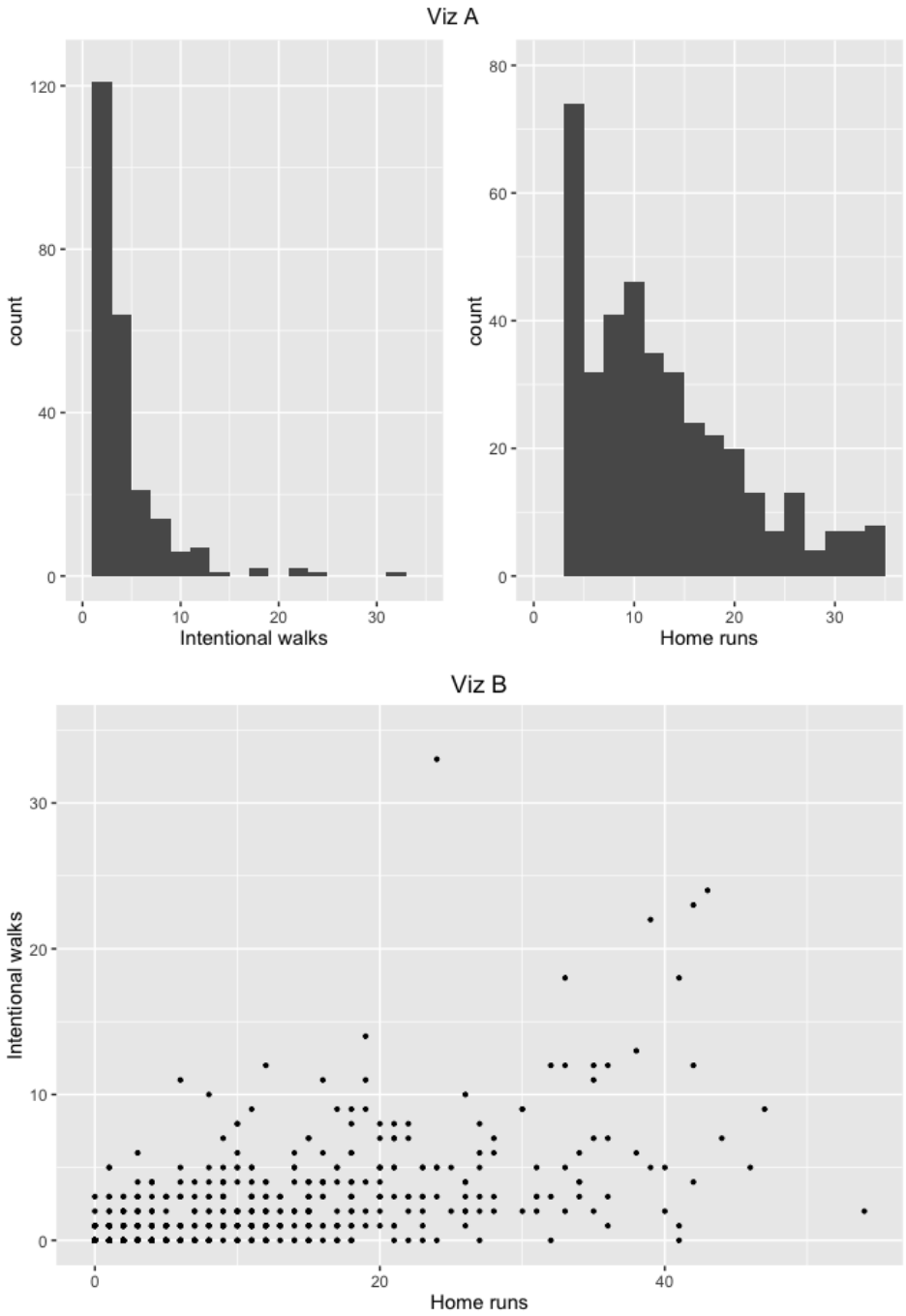
Introduction

This report presents an analysis of selected survey responses comparing various statistical questions between two visualizations (Viz A - histograms and Viz B - scatterplot, respectfully). The survey was conducted on the University of British Columbia 2016 MDS Cohort on November 16, 2016.

Survey Setup

The survey was administered using Google Forms and comprised the same twelve questions for both visualizations. The survey for Viz A can be found and the survey for Viz B can be found .

Viz A and Viz B are as follows:

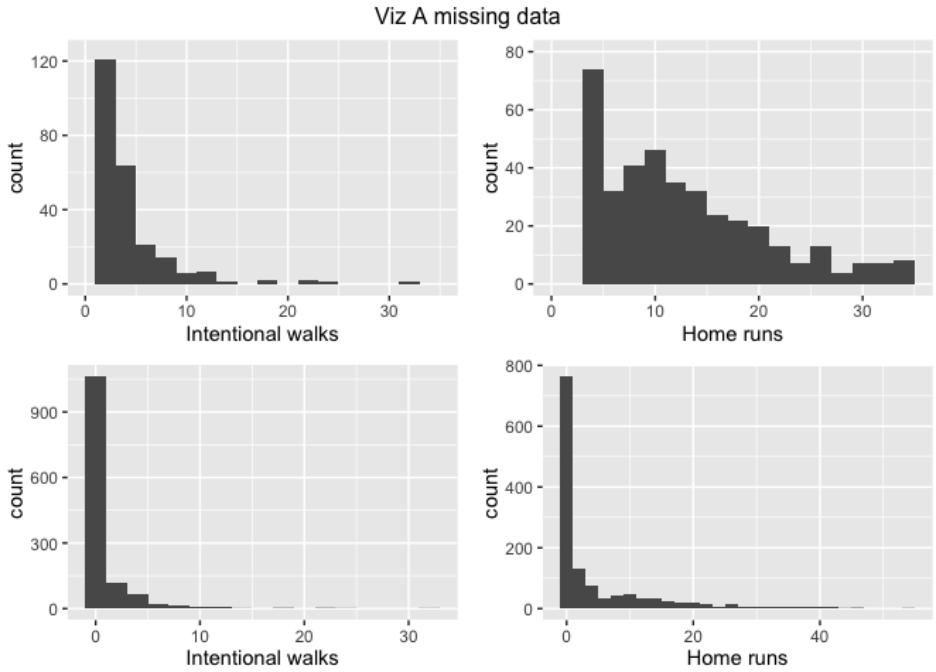


The following twelve questions were asked; however, this report only considers results that are statistically significant or of interest to visualization (bold highlighted).

1. **This visualization makes it easy to see the distribution of the data**
2. This visualization makes it easy to identify outliers
3. This visualization makes allows you to easily identify trends in the data
4. From this visualization, estimate the average number of intentional walks for a player
5. **From this visualization, estimate the distribution of the intentional walks**
6. **From this visualization, estimate the distribution of the home runs**
7. **From this visualization, estimate the number of intentional walks that are outliers**
8. From this visualization, estimate the number of home runs that are outliers
9. From this visualization, estimate the most frequent number of intentional walks observed
10. From this visualization, estimate the most frequent number of home runs observed
11. From this visualization, estimate the smallest number of intentional walks observed for any player
12. From this visualization, estimate the smallest number of home runs observed for any player

Potential Inaccuracies

The visualizations provided by the MDS program for use in the survey are potentially flawed. Viz (A) removes a large amount of data such that it makes it impossible to determine an accurate response to the question and which may have influenced the results of several of the questions (q7, q8, q9, q11, q12 in particular). This is due to the use of `xlim() and ylim()` in R which remove data outside of the range, in particular any bar that touches the limit. The visualization below shows the issue. It's unclear how this may affect some of the questions, however it likely influenced the results of all questions around outliers and anything asking for the smallest number (since the smallest number is removed from the histogram in both instances). Over 900 data points are removed from the histogram dataset for the "0" bar, which significantly changes the outlier prediction.



Data Preparation

The data was provided in .CSV format and contained some errors. The google sheet did not have any form validation resulting in character responses, non-character entries, and other such artifacts. Any data not strictly conforming to type numeric for any numeric question (everything but q5, q6) was converted to NA.

Approach

The questions were separated into three groups, with the following statistical test used to accept or reject the null hypothesis:

1. Ranked ordinal data (q1, q2, q3) - Wilcoxin Rank Sum Test
2. Categorical data (q5, q6) - Fisher Exact Test
3. Continuous data with one categorical variable (q4, q7 to q12) - T-test

A combination of one-sided and/or two-sided tests were run depending on the question.

Results

The full workfile for the exploratory data analysis can be found . The following two tables summarize the uncorrect p-value results:

question	method	p_value	null
q1	MW	0.000087	REJECT
q2	MW	0.004260	REJECT
q3	MW	0.460000	FDR
q4	Fisher	0.168600	FDR

question	method	p_value	null
q5	Fisher	0.454800	FDR
q6	t	0.646900	FDR
q7	t	0.318900	FDR
q8	multiple	0.011100	REJECT
q9	t	0.199600	FDR
q10	t	0.001086	REJECT
q11	t	0.021680	REJECT
q12	t	0.002474	REJECT

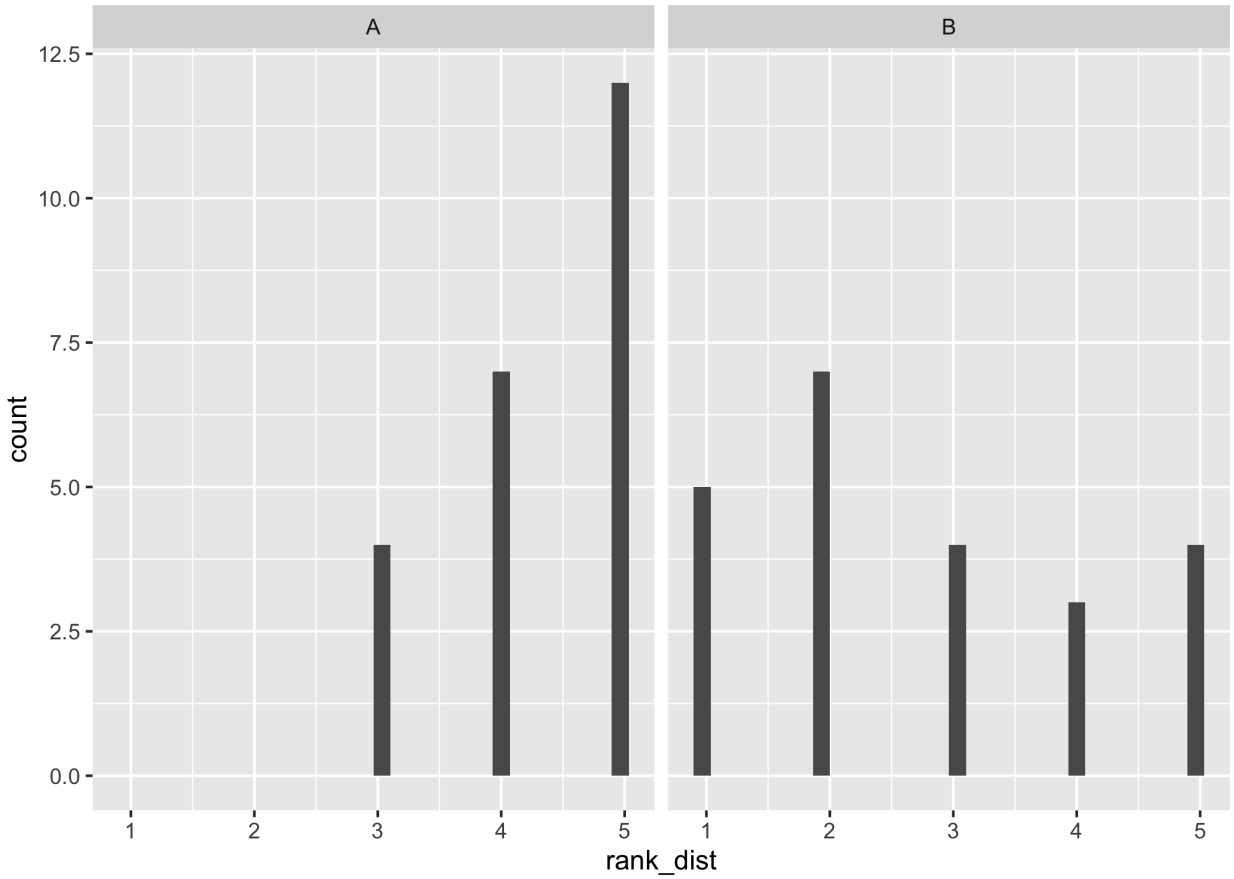
After controlling the False Discovery Rate at 5%, the following remain significant:

question	method	p_value	null
q1	MW	0.0010438	REJECT
q2	MW	0.0127800	REJECT
q8	multiple	0.0266400	REJECT
q10	t	0.0065160	REJECT
q11	t	0.0433600	REJECT
q12	t	0.0098960	REJECT

Discussion

Question 1

Question 1 asked for a score between 1 (strongly disagree) to 5 (strongly agree) for how easy the visualization showed the distribution. The results are shown below.

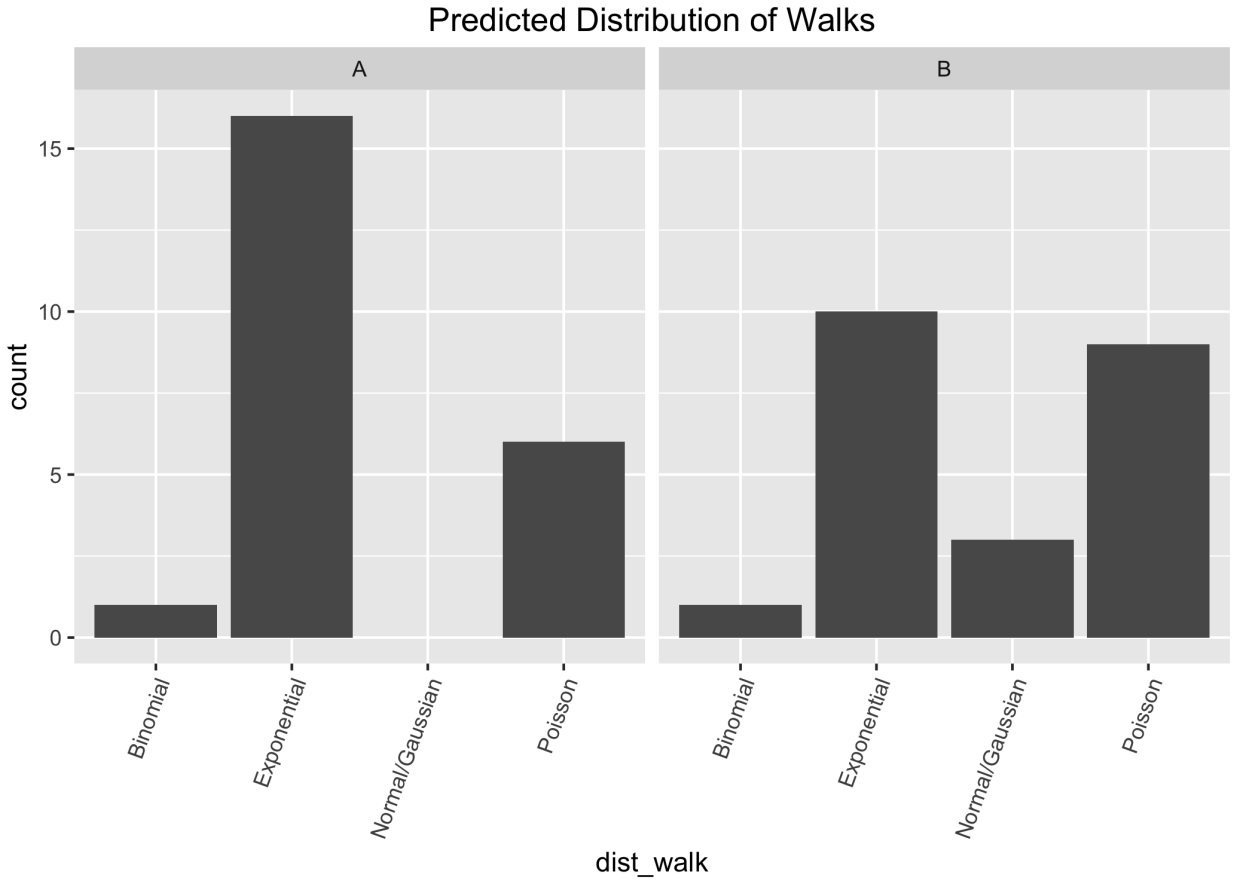


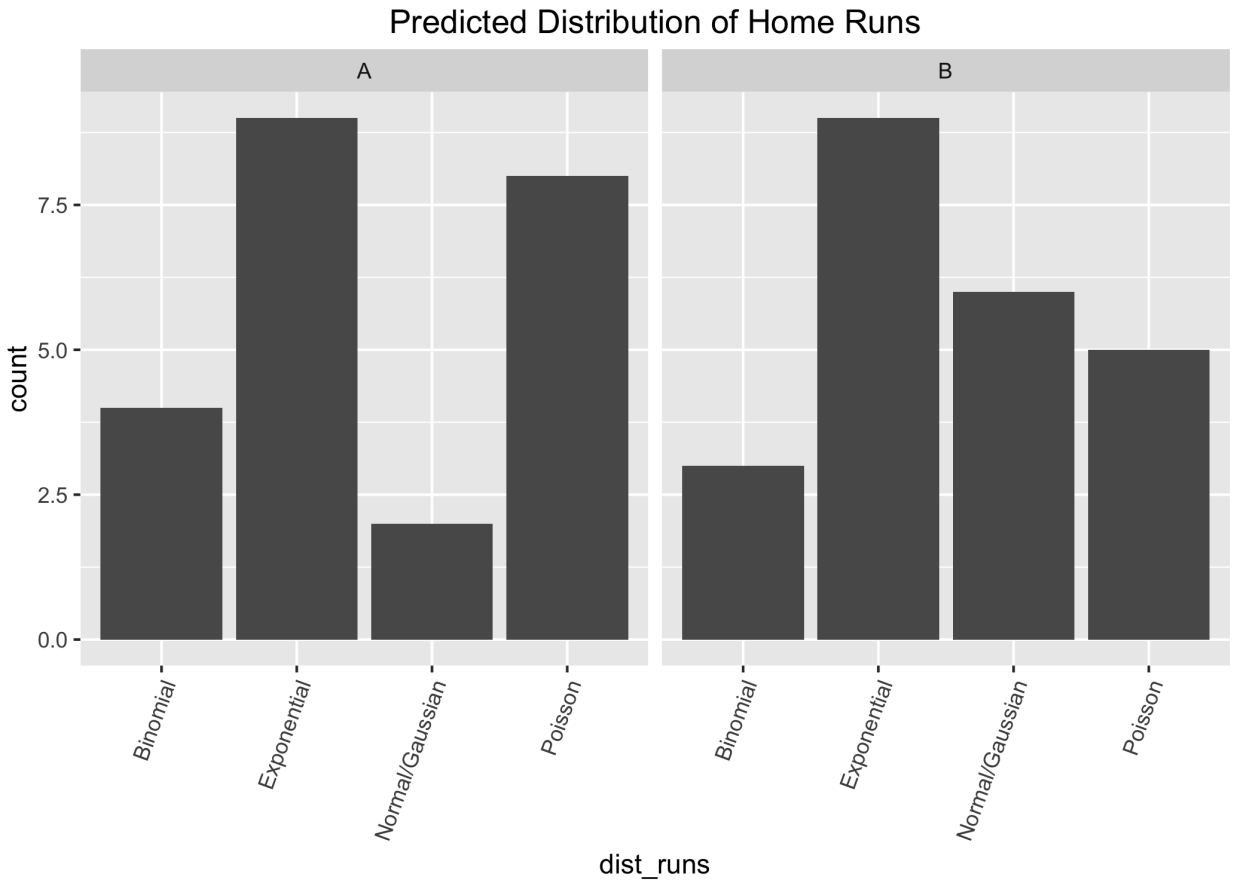
There is a clear preference for visualization A. A one-sided t-test with *Ho*: The two rank populations are equal (dervived from the same population with equal means). *H1*: Viz A has a higher median rank (positive shift in distribution) over viz B. We reject the null in this case, with a corrected p-value of 0.001.

This agrees with visual theory which indicates that histograms excel at displaying the underlying distribution of the data.

Question 5 and 6

Both Question 5 and 6 asked the participants to estimate the distribution of the sample population for walks and home runs respectively). The results are shown below.





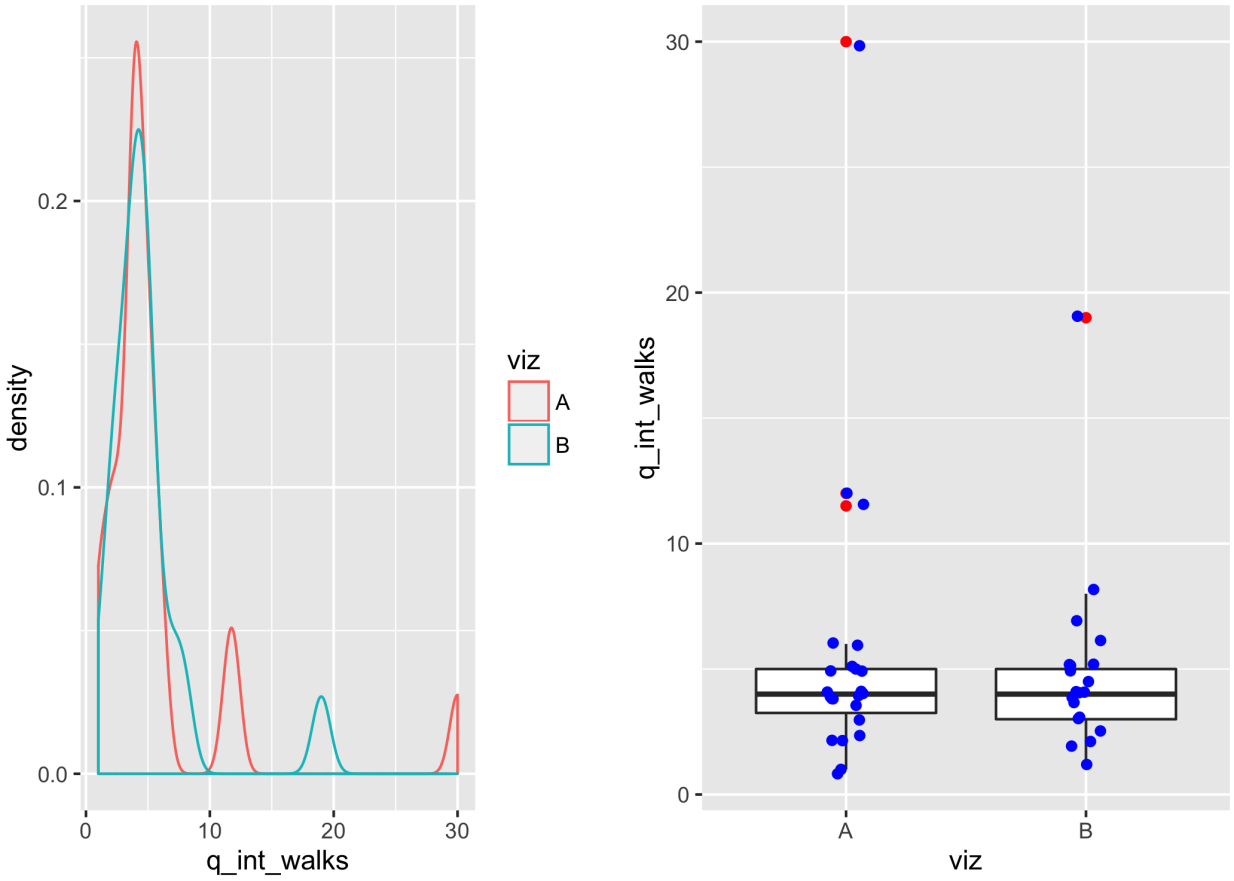
Expontential and Poisson were selected more frequently than the remaining options. Viz A more strongly suggests a Expontential distribution for intentional walks. For home runs, it is more difficult with Poission and Exponential being nearly equal and a larger showing for Normal.

A fisher exact test was used to test *Ho*: the relative proportions of the selected distributions is the same regardless of visualization and *H1*: The relative proportions of the selected distributions is not the same depending on the visualization. We fail to reject the null in this case, with a (uncorrected) p-value of 0.454 and 0.65 for Q5 and Q6 respectively.

The most selected distribution for both visualizations is the correct one (exponential); however, it is surprising that the scatter plot performed equally well. This could be due to the participants having undertaken Viz A first, and merely copying their answer for Viz B since they have already seen the question (and would be unlikely to change their answer having seen the visualization A which is superior for displaying the underlying distribution of a sample population). The distribution of home runs may indicate that MDS students are just poor at describing underlying distributions regardless of which visualization is selected. It should be noted that around half the survey participants did select the more correct answers (here considered to be Exponential or Poisson).

Question 7

This question focused on estimating the average number of intentional walks that are outliers. The survey responses are shown below:

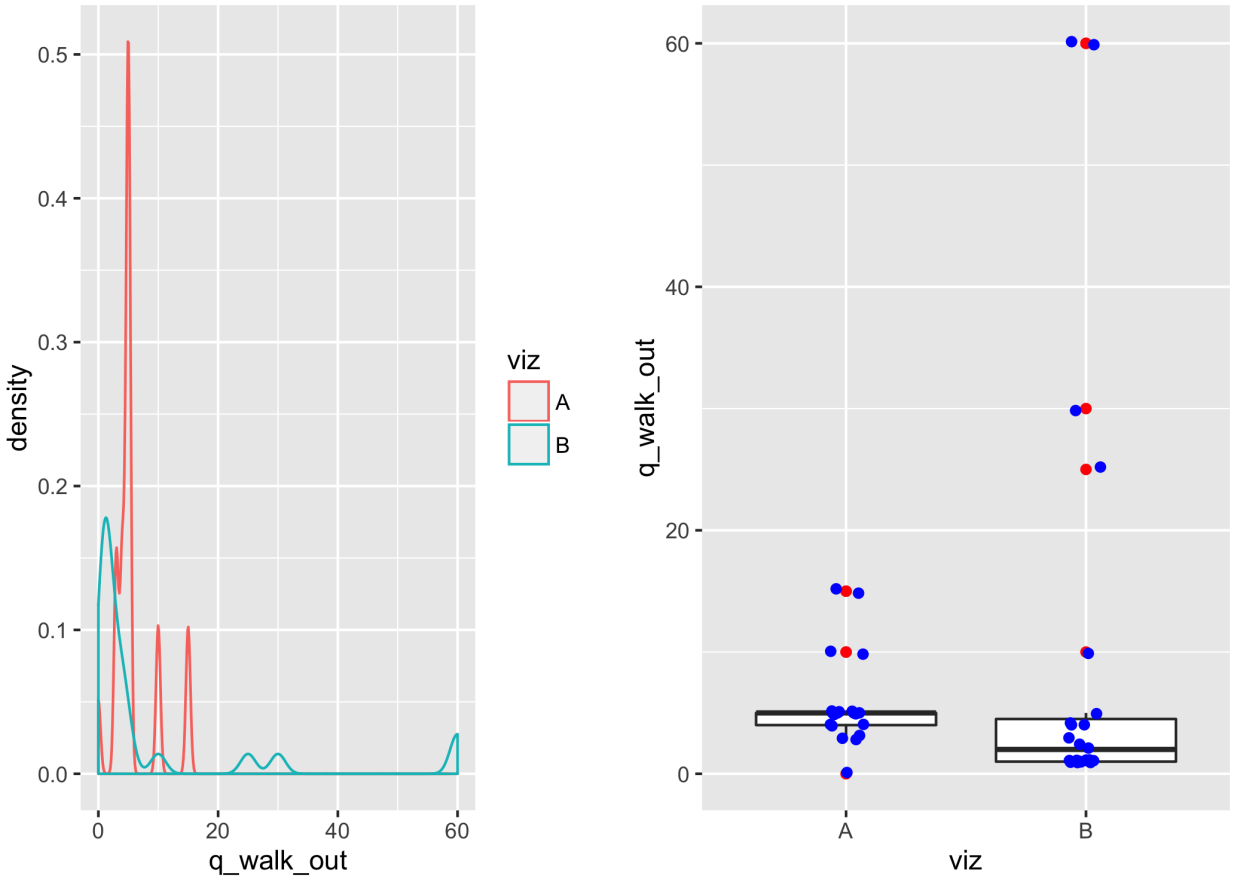


Visualization A had a mean of 5.76 compared to a mean of 9.5 for visualization B. A two-sided t-test failed to reject the null (*Ho*) that the average # of intentional walks that are outliers is the same between viz A and B.

In the actual underlying data, there is 175 actual outliers (the boundary was calculated as the 0.75 quantile + the IRQ). Even if Viz B had performed a little better such that it was statistically difference than A, it still would have had abysmal performace compared to the actual outliers.

The histogram performed poorly because approximately 900 datapoints around the zero bar are removed, which makes it look like there is significantly less outliers than there actually are. Scatter plots are supposed to perform better for outlier detection according to theory; however, that is not the case for out dataset. There are 900 points that plot on top of each other near x=0, y=0, and a significant cluster of data at the bottom left corner of the graph, which makes it difficult to visually find a boundary where the outliers may begin. If the data was less clustered, or a more linear trend was easily identifiable, outlier detection would have been easier in the scatterplot. As such, in contradiction of theory, I believe the proper histogram (see: Potential Innacuracies) is actually better at predicting the outliers.

This holds true for Question 8 as well, with the following result:

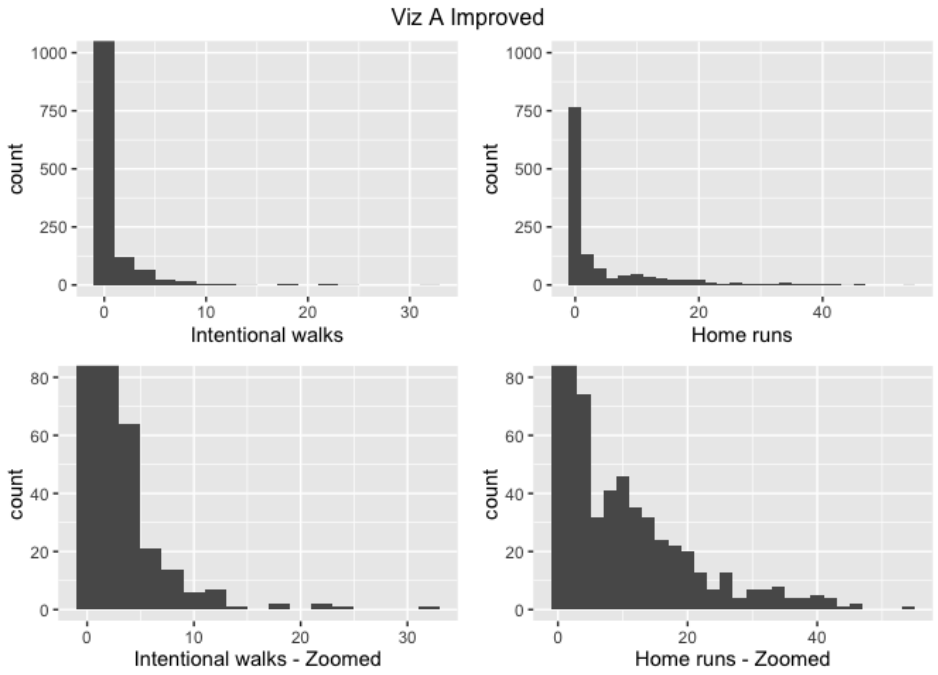


There are 190 outliers. Viz A and B have the same mean (at an alpha of 0.05); however, the histogram has a higher median (at an alpha of 0.05, by a one-sided t-test and still significant after p-value correction). It's odd that the histogram outpreformed the scatter plot which shoudl have been better at picking out outliers, likely for the same reasons pointed out above. However, the median outliers selected by the participans for Vis A is only 10, compared to 190 actual outliers.

Suggested Improvements to Viz A and Viz B

Viz A

Vis A could be improved by showing the full range of the data in one row, with a second row that shows a zoomed version of the data. Note that `xlim` and `ylim` should not be used but that `coord_cartesian` should be used instead. Suggested improvements shown below



Viz B

The points are converted to hexbins and a color scale is applied. This is a little hard to read since only two hexes actually get colored (the vast majority of the data near x=0, y=0). All the black hexes are basically outliers.