Derick Shi Xtern 2024 Data Science Report

This is a fairly simple report that will serve as the deliverable for this work assessment. The Jupyter notebook itself has several comments, but this will be a more compiled version of the findings.

One of the key insights that stood out to me from this data was that there wasn't really much correlation between the columns of data. Looking at the heatmap below, the value in each of the boxes represents the correlation coefficient between the category on the left and on the bottom. A value of 1 signifies extremely strong positive correlation while 0 represents no correlation and -1 represents strong negative correlation. Obviously the boxes that compare the same category to itself will have a value of 1, but the other categories don't really show much strong correlation. There is a 0.19 correlation between sponsorship correlation and chat activity but in general these numbers represent fairly weak relation to the categories.
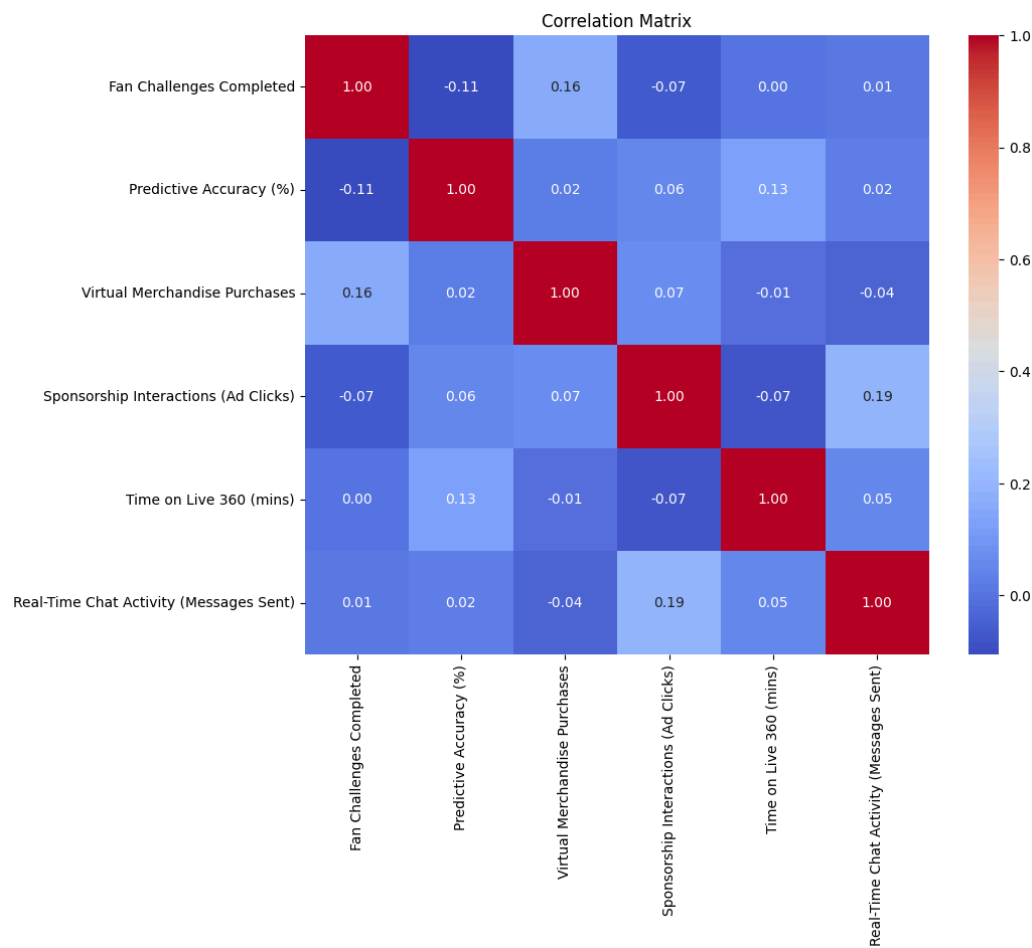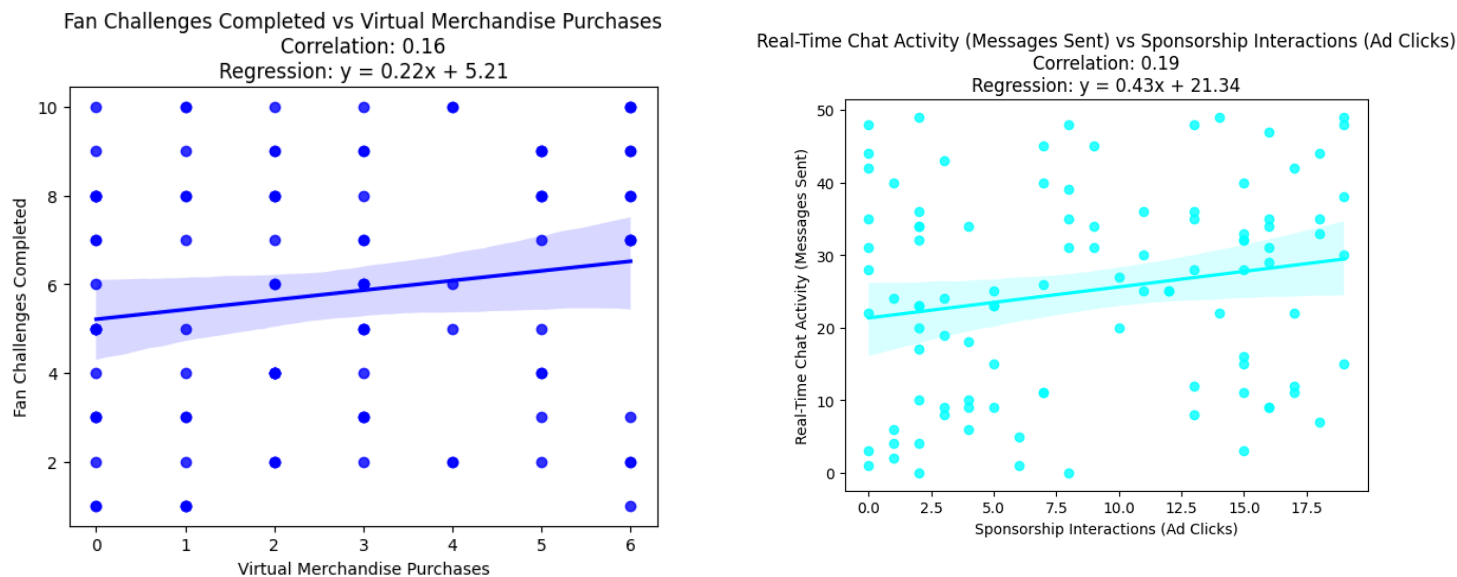
I also wondered whether or not such a lack of correlation could be due to the fact that users who really enjoy the Velocity event would interact with it more on every level as described in the data and people who don't wouldn't, but I don't think that seems to be the case either. There was only 1 fan that was much average for all 7 categories and only 27 fans that were above ½ times the mean value for all 7 categories so it seems like a fan's data is incredibly random based on this app.
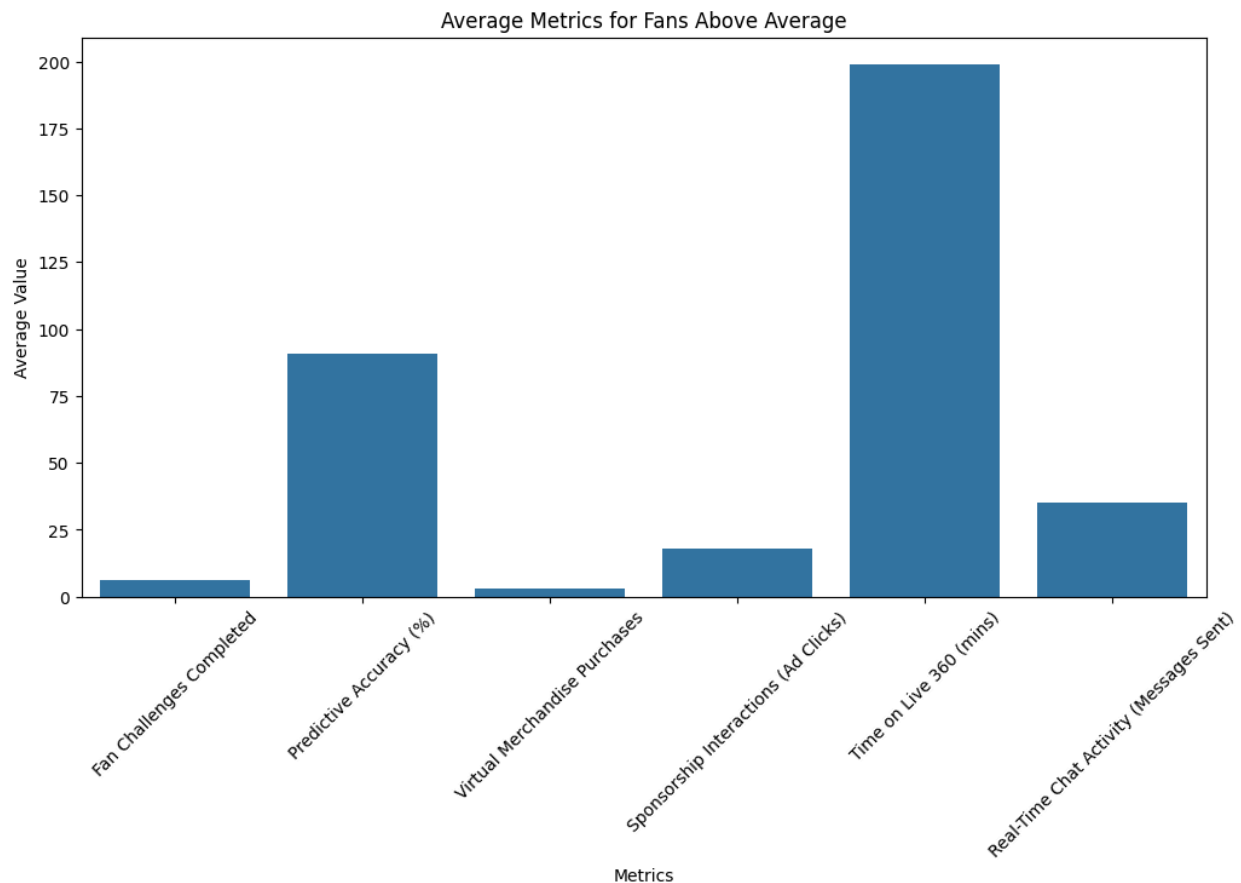

Besides having all of the data overall and a fan's data seem fairly uncorrelated, I was also curious to see if there were any outliers but there doesn't seem to be any. All of the data seemed to be within a standard IQR outlier test, and this was further supported by many box plots.

Box Plot for Fan Challenges Completed

Box Plot for Sponsorship Interactions (Ad Clicks)
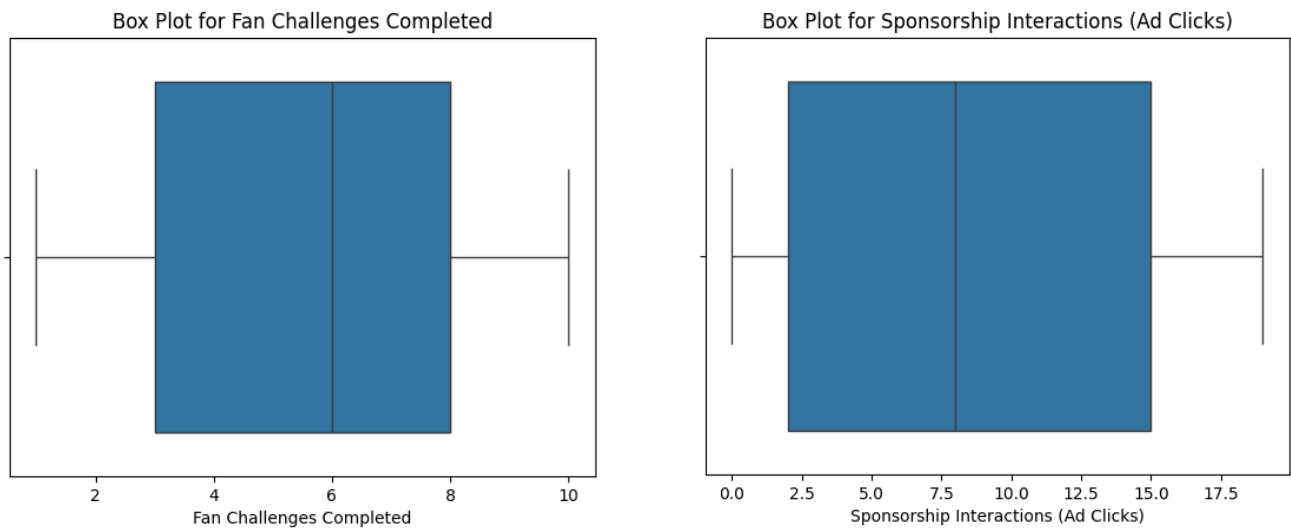
Regardless, the task asked for predictive models, so I ran some fairly standard machine learning models from the ski-kit learn library to develop a linear regression line with a 95% confidence level shaded. The strongest correlations were the same as described above between fan challenges completed versus virtual merchandise purchased and chat activity vs sponsorship interaction. It's interesting to note that the vast majority of the points lie outside the confidence interval, so these predictions are very weak, but it still describes a general formula to predict some of the variables.

Fan Challenges Completed vs Virtual Merchandise Purchases
Correlation: 0.16
Regression: $y = 0.22x + 5.21$

Real-Time Chat Activity (Messages Sent) vs Sponsorship Interactions (Ad Clicks)
Correlation: 0.19
Regression: $y = 0.43x + 21.34$

Derick Shi Xtern 2024 Data Science Report

I also clustered the 6 different columns into 3 weighted columns based on a user's rank in those 6 columns to normalize the data values. I hoped by moving the different categories into overarching categories of immersiveness, engagement, and monetization, that it would perhaps make the data more clear, but this approach also didn't do too much. For clarification, the numbers on the graph below represent the relative scale to the overall percentile of the user in that category so it only goes from 0 to 1. The most interesting was how if I made the outlier data bounds incredibly tight so that only 2/3s of the original fans remain in the data set, there was a bit stronger correlation between immersiveness and engagement/monetization than previously. This should show that the most average fan is affected a bit by these categories and that by improving immersiveness, it would in return affect the other categories.
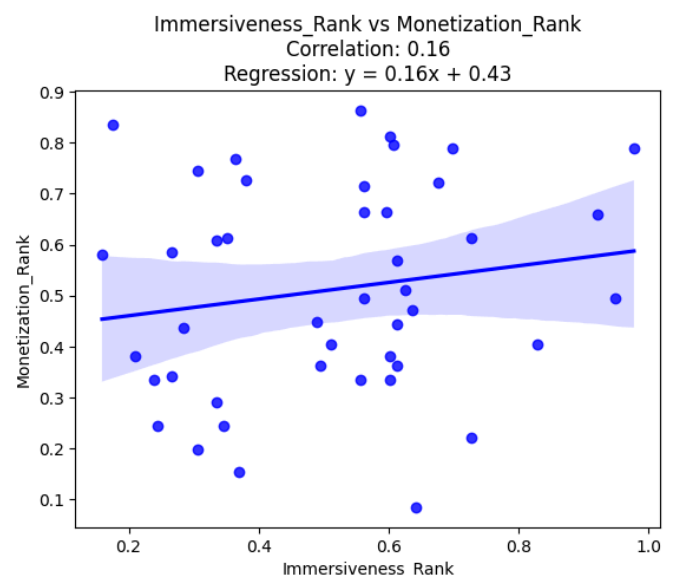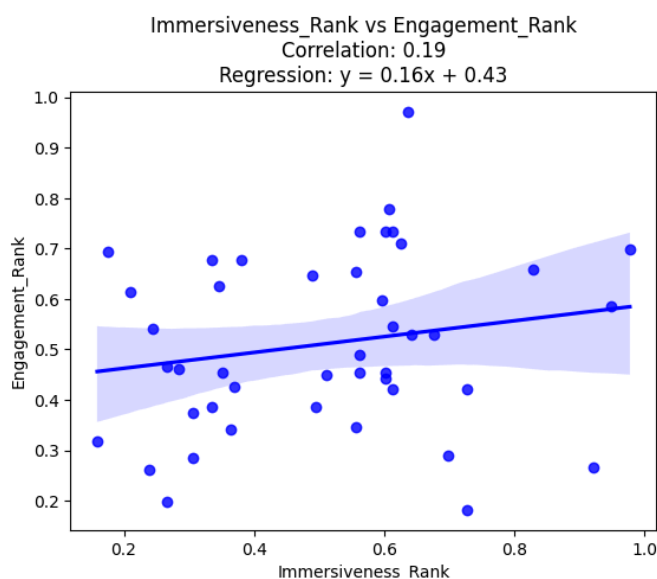
I also tried running different machine learning algorithms like decision trees and SVM, but they also didn't provide much insight and weren't very reliable.

Due to the randomness of all of these categories and lack of strong correlation between categories, it appears that incentives in integrating several of these categories to a fan could prove beneficial in increasing overall engagement with the app but also lead to greater monetization. Perhaps some of the user engagement games could give a lot of easy points if the user was immersed as they would know some helpful race information. Having engagement

game wins translate over to potential discounts on merchandise could also potentially drive greater integration between these categories and lead to a better fan experience overall. To exemplify these ideas, a potential new fan challenge that could provide beneficial for this app can be called "Speedstar" where a fan would guess the speed of different vehicles at different parts of the track and based on that receive points for merchandise. This would hopefully encourage greater fan immersion, as dedicated fans who open up the cockpit view of different cars would have an easier time completing this game, and this should hopefully also translate to a higher volume of sales as users should be able redeem their won points for merchandise discounts. If this resulted in the average fan challenges completed to just increase by 1, that would hopefully increase the average merchandise purchased amount by 0.22.