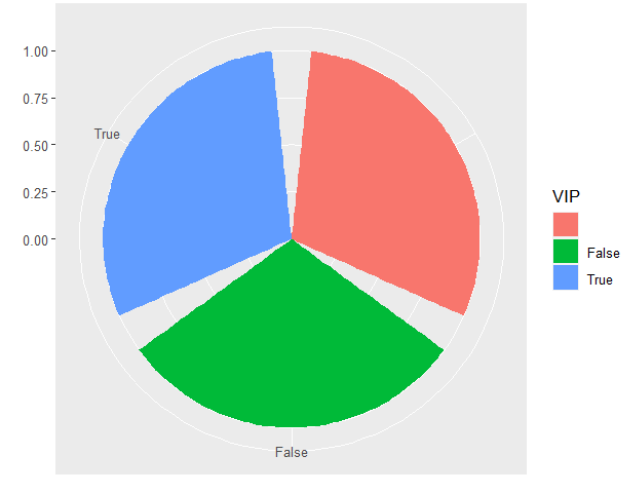
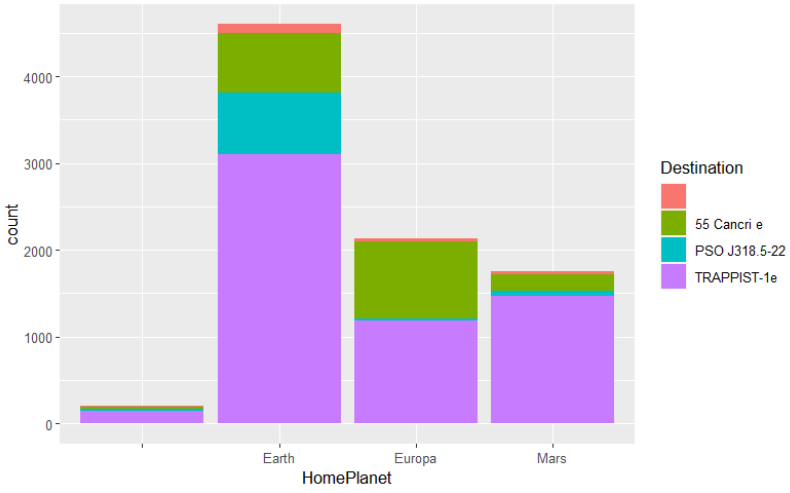
Data Analysis of Spaceship Titanic

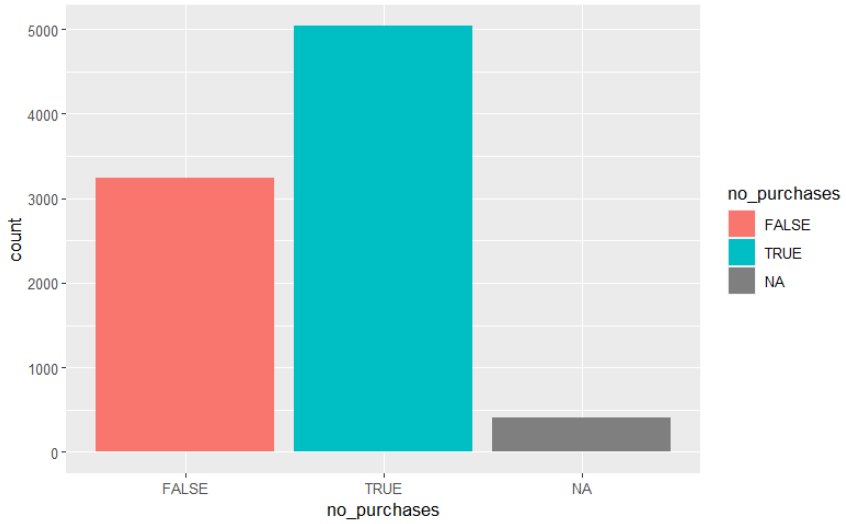
My initial query was to understand the distribution of VIP passengers and non-VIP passengers (fig.1). I was surprised to see how similarly balanced they were, as well as the large amount of Null values in the data. This led me to look more at the null values and specifically where null values in some columns had data in other columns. As a test set of data, it's simply interesting to have Null or empty values. However, if we were looking at a real data set, the amount of Null or empty values is pretty concerning and suggests a pretty significant issue with record keeping.



Next I was curious about which destination was the most popular amongst passengers - which turns out to be TRAPPIST-1e, followed by 55 Cancri e (fig.2). I compared it against the passenger's home planet - most of the passengers are from Earth.



I found it interesting how many passengers that were marked as having not been transported on the ship had transactions reported for room service, the food court, shopping mall, spa, or VR deck. If this data were a real reflection of an operating ship, I would be very concerned about how many purchases were being made by passengers not present. There is either a large amount of fraudulent transactions happening, or some serious gaps in recording. Either way, it's not what you want to see. It does however make for a more colorful graph.



I graphed purchases or transactions reported in food service, food court, shopping mall, spa, and VR deck - and grouped them by whether the passenger was reported to have been 'transported' or not (fig.3). The 'TRUE' bar is what they all should be - transported: false, no purchases: true. However there are just over 3,000 passengers who were not transported but did make purchases. Maybe it's recording all purchases ever made by that passenger, on the Titanic or not, but if this were real data from a real ship, we wouldn't want to see that.

Finally, because I was curious about how the randomized data played out, I wanted to see the age distribution of VIP passengers compared to non-VIP passengers (fig.4). The final graph shows that while the average age of VIP passengers is certainly older than the average age of non-VIP passengers, the oldest passengers on the ship are non-VIP. This actually came out pretty much the way I expected it to, though it is skewed a bit younger.

Overall it's an interesting set of data to play around with. I struggled a little to get creative with this project. I didn't have a clear idea of what I was curious about with this data when I started. Now that I've gotten to play around with it a bit more, I'm more curious about the values present for non-transported passengers. Knowing that it's randomly created, it explains the discrepancies, but otherwise it would be a really intriguing error in the data collection or passenger reporting.