**Inferential Statistics on the Capstone Project**

Restating the problem statement – using the data we got from the kaggle.com for Instacart Market basket we need to predict which product will be next in the user’s basket.

We did a small analysis on the Alcohol department.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Range of Hour** | **Afternoon** | **Evening** | **Morning** | **Night** |  |
| **weekday** |  |  |  |  | **Total** |
| **Friday** | 497 | 242 | 271 | 26 | **1036** |
| **Monday** | 439 | 196 | 309 | 14 | **958** |
| **Saturday** | 385 | 175 | 247 | 27 | **834** |
| **Sunday** | 355 | 138 | 171 | 13 | **677** |
| **Thursday** | 346 | 159 | 173 | 35 | **713** |
| **Tuesday** | 333 | 184 | 209 | 39 | **765** |
| **Wednesday** | 325 | 155 | 129 | 6 | **615** |
| **Total** | **2680** | **1249** | **1509** | **160** |  |

It appears from the above data that Friday afternoon most of the alcohol products are bought. Which make sense most people want to party on Friday night so they will buy alcohol by Friday Afternoon.

We make a Null hypothesis that buying more alcohol on Friday is just a coincidence. There is no direct relationship to it. We use Hacker’s statistics

To prove this hypothesis we will use bootstrap confidence interval method. I divided all the alcohol products that was ordered on Friday in 1 array and all the alcohol products that was ordered rest of the day in the 2nd array.

I merged these 2 arrays considering no difference between alcohol products bought on Friday or any other day. We will use Python library of ‘random’ to do that. We plotted Cumulative Distributive function and check how alcohol order is distributed over the week.

I got the 95% confidence interval with that it got proved that the above Null hypothesis is not correct.

I also used Permutation sampling and got the ‘p’ value to see how this Null hypothesis worked. It failed there also.

I concluded that Friday is the day when most number of alcohols products are bought.