Executive Summary

As soon as we realized Goldman Sachs was hosting a challenge, it got us interested. The problem looked open-ended enough to not restrict creativity while avoiding being vague or unclear.

We decided to analyze the dataset and realized it was all tacos and burritos and that got us even more interested. What data insights could we gather from something as trivial as those delicacies? Little did we know it would turn out to be an adventure.

# Plan of Action

We decided to get our hands dirty by solving the first question posed on Kaggle: what is the ratio of burritos to tacos for each city? We decided to iterate through all data values for all cities and count the number of instances of “burritos” or “tacos” in menu names and divide it, then store the results in a dictionary where the key is the city and the value is the ratio. Since we had several thousand cities, we decided to take random sample of size 15 and plot that in a scatter plot. We never realized how variable the ratio is! It ranged all the way from 0 to over 3 times as many burritos as tacos.

We tried to find correlations between several pairs of variables, including but not limited to cuisines, latitudes, longitudes, prices, etc. After not getting any clear direction or headway, we came up with an innovative idea—after an astute observation by us, we saw some of the restaurants had “takeout” or “delivery” in their description. We felt we could capitalize on that and create a collaborative platform with delivery companies such as Post Mates. Areas with a high density of takeout naturally would have a higher demand for delivery companies, giving Goldman Sachs the opportunity to gain secondhand profit. Meanwhile, we could point out the restaurants missing out on profit by not having a delivery service, in delivery-dense areas.

But none of those ideas worked. There weren’t enough companies who had “takeout” or “delivery” in their descriptions. We could find 4 for the former and 68 for the latter. This wasn’t nearly enough to draw stable conclusions.

# Ray of Hope

We didn’t lose hope though—we kept brainstorming, running regressions, and analyzing data to come up with an innovative idea to get a great insight into the data. Eventually, we realized since we were provided with a lot of categorical data, rather than encoding them and trying to perform predictions, we decided to embrace the categorical variables such as cuisines, menu names, descriptions, etc. and create word clouds for a particular type of restaurant.

First, we created a word cloud for higher priced restaurants, and we got an eye-opener. Most of the expensive restaurants had some form of “Southwestern Traditional American” in their cuisines. We decided to run that for the lower priced restaurants too, and then we realized the cheaper counterparts had a lot more diffused word cloud—with more variations in words, especially words like “contemporary” which stood in sharp contrast with traditional dining.

Finally, we decided to get some additional data from the complete dataset and obtained the working hours for 50 different restaurants chosen randomly. Then, we decided to find the cuisines of those restaurants open till late (after 10 pm) and got an insight we might be all familiar with—junk food restaurants like McDonald’s were open for much later than their healthier counterparts.

# Conclusion

We learned a lot from working for Goldman Sachs challenge—we learned before you make hypotheses, you run regressions.

We never thought looking at burritos and tacos data would take us on such an insightful journey. If a picture is equal a thousand words, then a dataset is equal to a million. Every dataset tells so many stories—stories meant to be found out. That is why we have data science.