Product Scientist take home

1. *How many leads are represented in this dataset? Describe both the assigned and unassigned populations. What is the average revenue of each group?*

    a. There are **77875 total leads** in the cleaned dataset. There were **77891 leads in the raw dataset**. 37064 were assigned to sales and 40811 were not assigned. The **mean revenue for leads that were assigned was 3240157 compared to 878927.8 for those not assigned**. The mean total_age of the account for leads that were assigned was 755 compared to 137 for those not assigned.

2. *What are the most important metrics to consider when answering the problem statement?*

    a. To answer the question - how much more did these leads spend because there was intervention. We cannot simply compare the two populations and compare the difference in revenue because they were not randomly assigned. To put it simply, the leads were assigned based on potential value, and we know there are differences in the populations (e.g. age). If we were designing this from scratch we could randomly assign leads and then test for differences in the means. Since we didn't do this, we cannot look at the difference between the means (3240157-878927.8) or the difference in total revenue and attribute it all to sales intervention.

    b. The question we are really interested in is if our assigned leads would have converted to revenue even without our intervention. This is a trickier problem to solve. We first need to consider the overall relationship between our data and revenue, and then we can test to see if sales intervention is a significant predictor of revenue.

    c. The main predictor we have here is **total_age**. I would first look at the correlation between total_age and revenue by developing a regression model. To see if sales is having an incremental increase, I would perform a stepwise multiple linear regression to see what additional variance in sales is explained once sales assignment is added to the model. I will explore if assignment as a categorical variable (0/1 not assigned/assigned) is more informative or of assign_days is more informative in predicting incremental increase.

3. *Analyze any existing relationship between account age and revenue.*

    a. There are several things that I might want to do before developing my regression models. One would be to scale my features using std or maxminscalar. I also might want to log transform my response variable (revenue) since I doesn't appear to be normally distributed. In the interest of time, I am not going to do those things now. I also am not sure about whether I want to include the data with revenue = 0. If I do include it, I should use zero inflated poisson regression. The revenue data has many zeros and regular MLR is not appropriate. I will use zero inflated poisson regression which is designed for this type of data.

My first step will be to develop a GLM (poisson regression model) using the python library Statsmodels and then I will develop zero inflated poission model for the entire dataset. My zero inflated poisson model didn't converge, so I ended up modeling only non-zero revenue values using a GLM with negative binomial link function.

b. **The coefficient for total age is 0.0012 and it is significant.** How do we interpret this? **This means that for accounts that had revenue, for every additional account day we are likely to have an increase of 0.0012 cents of revenue**.

4. *What is the incremental value of assigning a lead to the sales team?*

    a. Sales intervention/assignment has a strong incremental increase on revenue. The coefficient for 'assigned' is 0.90, this means that **for accounts that had revenue, $0.90 of every dollar of revenue is due to sales intervention**.

5. *Investigate the data however you wish and discuss any interesting insights you can find in the data.*
    a. I would like to look at whether the day of the year of first revenue is a predictor of revenue (maybe customers are only buying at a certain time of the year)? I will use datetime to convert first_revenue_date to day of year, then I will rerun my GLM with this added predictor, day_of_year.
    b. After rerunning the GLM, we see that day_of_year is a significant predictor in our model with coefficient of 0.0007. There is a slight improvement in this model as measured by a decrease in the log likelihood.