Grace Mongan
IS 497
Project
April 26, 2021

1. For my dataset selection, I want to use a premade dataset that I've seen before online which collected data about Cartier jewelry and products. I love hearing about prices and quality of expensive jewelry out of pure curiosity. If I am unable to locate this exact dataset I want to find something similar because it's a topic I'm interested in and I like looking at data containing prices as well as diamond quality amounts of data.
2. The problem I want to solve is trying to clean data more using Open Refine. As I look forward to a career potentially in Data Analytics I think about the different aspects of work I will have to do and cleaning data is a part that I don't have a lot of experience with. I want to take this opportunity to see how I can clean quantitative data and how to do it in a way that makes sense. Essentially, I want to see what the data cleaning process could look like.
3. The dataset I am planning on using is the Cartier Jewelry Catalog dataset that is posted on Kaggle. The author is Marcelo Pesse and I actually found this from Stat 107 discussing the different data sets they have on Kaggle. The information was also created from the Cartier in early 2020 so the data is most likely still accurate. On Kaggle it says," This dataset contains some info on the pieces scraped from the Cartier website on 04/2020.They have ref. number, price, some info on composition and some general info on the pieces available on the catalog." I decided to go down a deep rabbit hole and start looking at all of the different sets they have and why they were created. I really like seeing what other people have collected data on and why they have considered them important.

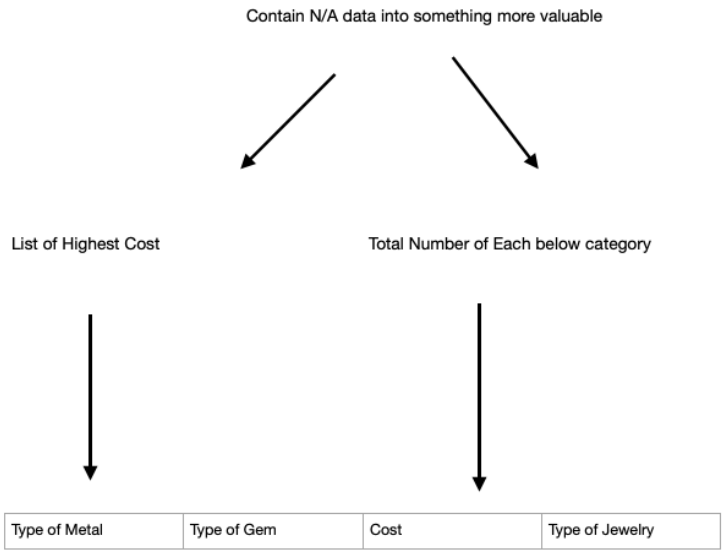
https://www.kaggle.com/marcelopesse/cartier-jewelry-catalog

I decided this is interesting because I love jewelry and looking at peoples different jewelry online. I especially enjoy gold jewelry which Cartier specializes in and has become so popular that I love to see what the data has been collected on it. This is something I want to explore because as time has progressed jewelry has so many different brands and a saturated market. I want to especially look at the price to quality of gold and diamond ratio. I am mainly curious to see these categories and how they are affected by the different types of jewelry they are. We are also able to look at the different types of gems they contain as well as the different types of metals. I think I want to split up the information and try to clean based on gold vs silver as well as rings vs bracelets vs earrings to further the exploration it contains. This data comes straight from the Cartier jewelry catalog which actually is the true amounts and exact quantity of elements inside each piece of jewelry. We can see several distributions here and also can tell how different the quality of jewelry is related to the price and materials included.

Overall, I think this is interesting because I find it interesting. I think data and datasets may seem a bit boring to others and too black and white with just pure numbers or random content being created that may not be as relevant to the average person and some

just might not care. I for example do not really care about a car part dataset but to me who is a huge jewelry fan is obsessed with a dataset related to my own area of interest. I of course do not expect others to think that this dataset is the most interesting to them because every dataset topic is different and geared to different people who enjoy gathering that information. Data is mainly interesting to the topic it is referring to, not that actual numbers. That is why I decided to choose this dataset because it is something I find to be interesting and would be relevant to myself and I want to try a new skill on something I enjoy. Trying out data cleaning more is a bit intimidating and I find kinda intimidating to me because I do not have much experience. However, I think a dataset such as this one with so much data of both kinds of quantitative and qualitative, makes me think that I will have a more interesting experience of beginning this data cleaning journey. I also think since I am interested in the topic of jewelry, doing a more intimidating task on something I find interesting will be a way to motivate myself to try new things and gain more confidence.

4.



My final project:

I first began with the raw dataset that I found from Kaggle. It is available on the public Kaggle page and I have the citation listed down below.
The only numerical data was the price, so when I made my histogram the y-axis doesn't make much sense. However, it did show the vast majority of its jewelry line is between 0-$50,000 and it drops off drastically after $100,000. This isn't very useful to me because there is no way I

would afford that so I dug deeper into the mean price. After realizing $27,057 is the average price and thinking that seems super high and almost unlikely for such a popular brand to be that costly, I did a two sided z-test of the price column and the average. My results told me that I was incorrect and the average is in fact very probable. The 95% confidence interval also crushed my dreams saying that we are 95% certain the true mean lies between $24,397 and $29,717. So, I decided with my whopping $2,500 saved in my bank account I would create a new column to see what percent of my entire savings could buy each item in column 'percent_of_my_money'. I ran another z-test against 100% of my savings and a 95% confidence interval that showed we are 95% confident that it would take at least 9.19 times - 12.45 times of my savings to buy anything Cartier. This equates to $22,975- $31,125 which makes sense due to my confidence interval of the mean price.

I did the same calculations to the second and tried to find the comparison between the two different data sets. We can see that the cleaned data was able to produce a smaller mean which shows that the overall outliers of the original data set were taken out and we have a more accurate mean of the true average of the amount of money it is to buy Cartier jewelry in general. I believe these outliers are mainly large purchases such as the most expensive season collection piece of jewelry that is $30,000 compared to the average of $10,000 which was shown through the data visualization using the new dataset made through Open Refine. We can see these changes and the benefit of data cleaning in this sense and can better understand our data through visualizing the data and being able to compare the previous data findings to the new data findings which is cleaned.