Title Slide

Include the name of the Project and Group Members

* Wine tasting review
* Alex, Arlen, Lilit, Anna

Motivation & Summary Slide

Define the core message or hypothesis of your project.

* We will be analyzing the data to find possible correlations between price and the # of points for wines across the world. Current hypothesis: Higher wine prices will have better ratings.
* Going further into the data, we plan to analyze different countries or regions to review popularity/price/types. Current hypothesis: US holds majority of the wine worldwide and holds higher overall ratings.
* Looking at variety types of wines, we want to compare the average year to to the price of the wine. Current hypothesis: the more a wine is ages (the older the year) the more expensive the wine will be

Describe the questions you asked, and why you asked them

* Is there a correlation between wine price points and their overall ratings?
* Which countries produce the most wine that is up for evaluation?
* Are certain varieties of wine reviewed higher than others? (i.e. does Pinot Noir have generally higher rating than a Cabernet?)
* Does the year a wine was made (aging process) play any factor in the price or rating of a wine?

Describe whether you were able to answer these questions to your satisfaction, and briefly summarize your findings

* We were able to find a very strong correlation among price and points.
* We were surprised to see that the aging process of wine didn’t play any factor in price/points of a wine - contradictory to popular belief that the longer a wine is aged the better it is.
* Our evaluation on the top varieties and reviewers further strengthened our argument that price and points have a strong correlation

Questions & Data

Elaborate on the questions you asked, describing what kinds of data you needed to answer them, and where you found it

Data Cleanup & Exploration

* Possibilities:
  + Compare expert ratings and user ratings
* Data cleaning:
  + Two datasets found, one contains more variables. For questions related to variables that are only in one dataset, we will use a subset of the data.

Describe the exploration and cleanup process

* We went through different datasets within kaggle.com and found an interesting dataset regarding wine reviews.
* Cleanup process: the dataset had two csv files, one with approximately 130K rows, and another with 150K rows. We combined the two datasets into one dataframe using pandas.append function.
* During this process we noticed that the two csv files have different # of columns. While majority of the columns match, there is useful information within the file with more columns that we decided we could use for further analysis. To use all applicable data, we will be using two DataFrames, one with all the data from both files, and another with just the data from the 130K row csv (with more granular data).

Discuss insights you had while exploring the data that you didn't anticipate

* We didn’t expect the aging process to be so arbitrary and have no correlation to the price or the point of a wine.

Discuss any problems that arose after exploring the data, and how you resolved them

* While exploring the data one issue was the fact that the two csv files didn’t have entirely matching rows so we decided to use the appended csv and the smaller one so that we can utilize the most of the possible data.
* We wanted to compared taster rating vs user ratings to see if any average person would rate wines similarly to tasters, but we couldn’t find a public dataset that contained this data.

Present and discuss interesting figures developed during exploration, ideally with the help of Jupyter Notebook

Data Analysis

Discuss the steps you took to analyze the data and answer each question you asked in your proposal

* Download data from Kaggle.
* Merge 2 datasets on common variables.
* Clean up data.
* Explore data by plotting and looking into subsets.

Present and discuss interesting figures developed during analysis, ideally with the help of Jupyter Notebook

Discussion

Discuss your findings. Did you find what you expected to find? If not, why not? What inferences or general conclusions can you draw from your analysis?

* We found that price is the highest correlating factor to a point a wine would receive.
* We did not expect to see that older wines did not have higher prices and points in association to this. We assumed that since aging a wine takes more effort and resources, it would ultimately be more expensive. Also, why would someone take the effort to age a wine if it isn’t even going to be good?
* General varieties have price points that correlate to the rating they received.
  + ULTIMATELY price of a wine is what matters most.

Post Mortem

Discuss any difficulties that arose, and how you dealt with them

Discuss any additional questions that came up, but which you didn't have time to answer: What would you research next, if you had two more weeks?

* I would search further to find user ratings.

Questions

Open-floor Q&A with the audience