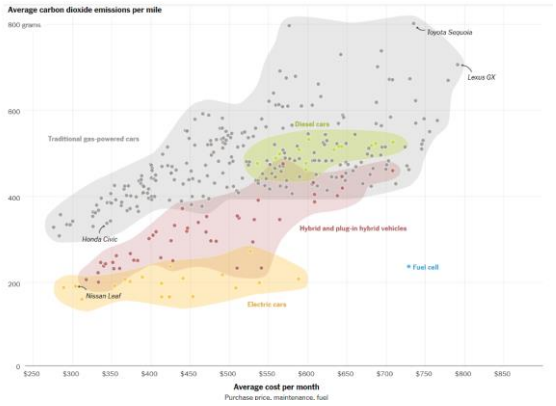
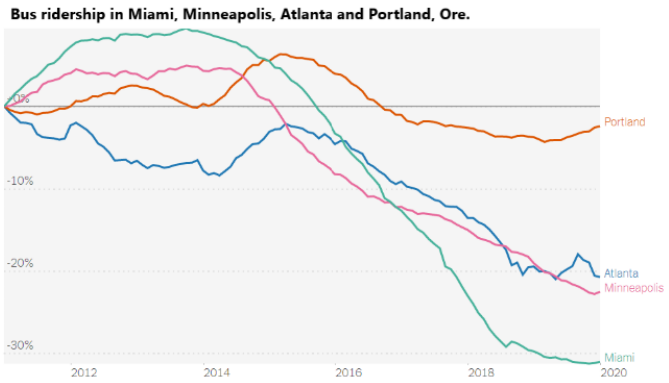
CS260 SPRING 2023 PROJECT

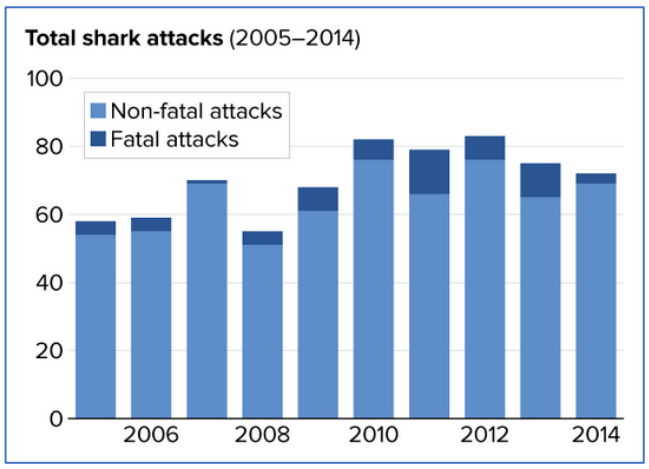
In this class, you will learn to create the following visualizations (see below for examples), all of which you have certainly seen in your lifetime.

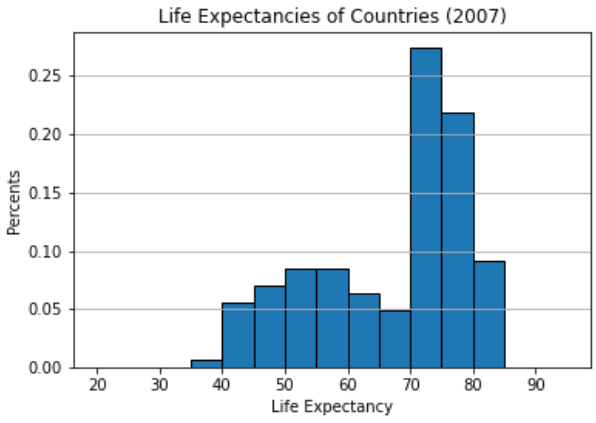
These different types of visualizations help us learn about data. Some plots are used for numerical quantities (time, height, number of times a year someone gets on a plane) versus categorical qualities (hair color, favorite ice cream flavor, type of car someone owns such as sedan, SUV, etc).

- Scatter plots: Given two numerical variables, scatter plots help us answer questions like “Are the variables X and Y related/associated?” or “As X increases, does Y tend to increase too?”
- Line graphs: Analyzes trends across time or other numeric values.
- Histograms: Helps us understand how numerical data is distributed across multiple bins.
- Bar Charts: Helps us understand how categorical data is distributed across multiple categories.









For the final project, your goal is to find an interesting dataset that will lead to interesting visualizations. I will help you load your data into a sql lite database so that you can query it. Then you can use the Python we learn to make the pictures.

PHASE 1 Directions

With your group, find a dataset that meets all of the requirements listed on page 2-3 of this document.

Once you decide, post the link to your data on the discussion board so that no one else uses the same data set.

Copy and paste the template on the last 2 pages of this document into a word doc and submit a word or pdf doc called "Proposal_Initials_Initials" it by the due date. Example for a team of 2 people: Proposal_KR_PG

DATA SET REQUIREMENTS

Here are some rules you must follow when choosing the dataset.

- **NO LAME PROJECTS:**
    - Find something that interests you, truly! FYI: For those of you who are completing the data analytics minor, this project will be part of your portfolio in a future class. So make it as nice as you can.
    - No Pokémon data sets. Sorry. I've seen way too many of these.
    - If you are using a data set about video games, you need to ask me or email me for permission. Why? I've seen a lot of these too, and there's a few where the data is just yucky. So I'd like to prevent you from going down that route.
    - Note: You cannot use a dataset which someone else claimed. Datasets are first-come/first-serve. To stake your claim, post a link to your dataset on the designated area of the discussion board so others can see it.
    - You need to be able download the data as a CSV or an excel file. If you can't do this easily, you should come see me asap so I can help.
- **THE DATA SHOULD BE LEGIT/BELIEVABLY VALID:**
    - The data set must come from a legit website that conceivably has reasonable/believable data. How can you verify this? Well, the website from which you acquire the data should state when and how the data was collected or should refer to a secondary legit website that has these details.
    - The website should explain every variable/column within the dataset. Why? So that you know what you're analyzing.
    - This website should not include a bunch of visualizations regarding the data. Why? Because your job will be to think of reasonable questions to ask that can be answered with pictures, not to just recreate pictures. So any visualization shown on the website ***cannot*** be used in your final project.

- **NUMBER OF ROWS:**
    - The data set should have at least 500 rows.
    - Also, I'd suggest the data have less than 1 million rows. However, if you find a dataset you like that has more than 1 million rows, then talk to me about it and we'll decide together if it's a good pick.
- **NUMBER OF VARIABLES/COLUMNS:**
    - The data set must have at least 12 columns of varied variable types.
    - Some columns should be numerical (with a mix of continuous and discrete data) and some should be categorical. For example, a data set about flights might involve the length of flight (numerical - continuous), the number of times it flies a week (numerical – discrete), and the name of the airport of departure/arrival (categorical).
    - Note: Sometimes students incorporate columns that are the same except for units. For example, I've seen folks have a column that is for the population per 100000 and another column for population per 1000. *These do not count as different columns/variables.*
- **A COLUMN THAT WILL ALLOW YOU TO COMPARE DISTRIBUTIONS OF "TYPES":**

  Every project needs to compare types/subgroups. For example, at the link below, we show how diamond depths compare between ideal/fair/good diamonds. This means there needs to be a column with Ideal/Fair/Good in the dataset as well as a column for the depth of diamond.
    - https://www.machinelearningplus.com/wp-content/uploads/2019/02/Histogram_12_0-min.png
- **A COLUMN THAT CAN BE 'JOINED ON':**

  Every project will have to involve a "join" of some sort. This means there have to be a column in your dataset where it would be feasible/possible to join your data with a column from another table/datasetto answer a question.
    - Example: In the Big Foot data set, there was a column for the date of each sighting. From this column, students generate a column with just the year of the sighting. The team then used another unrelated table that had a column for the year as well as a column for the population of the USA in that year. The team then paired up/joined the data from the first table with the second table to generate a table that contained the year, the number of sightings in that year, and the population of the USA in that year.
    - Example 2: Say an Air B&B dataset/table contains a column for the city of the rental. Say a second table listed major cities in the USA with their latitudes/longitudes. Then a united table could be created that pairs the number of Air B&B's in each city with the city's longitude/latitude. From this new table a bubble map like the one here could be made.

CS260 SP 23 Final Project Phase 1 Template

NAMES OF TEAM MEMBERS:

- Brendan Hasara, Matt Chylack, Travis Kerr

DATA DETAILS:

- LINK TO THE DATA:
  - https://data.world/etocco/nba-team-stats

- EXPLAIN HOW/WHEN THE DATA WAS COLLECTED
  - The data was collected on Monday Feb. 6th by Google searching for datasets that included NBA statistics from previous seasons of the NBA.

- EXPLAIN WHY YOU BELIEVE THE DATA IS VALID/LEGIT/BELIEVABLE
  - I believe that this data is valid and reliable because the owner of this dataset used a secondary legit website where he gathered all of his data from. It includes a dictionary describing each of the columns within the dataset and does not show any visualizations as well.

- EXCEL/CSV PLAN: Are you attaching the CSV with this submission? If not, what is your plan to download it or get access to it? Did you discuss this plan with me?
  - The link we have attached already gives access to its CSV file for a very easy use of the dataset.

- HOW MANY COLUMNS AND ROWS ARE IN YOUR DATA:
  - Columns: 22
  - Rows: 725

- FOR EACH COLUMN, GIVE THE NAME, DESCRIBE WHAT THAT FIELD MEANS, GIVE THE TYPE OF THE VARIABLE AS (so Continuous/Discrete/Categorical) AND THE SCHEMA DATA TYPE (so text/real/integer/blob):
  - no - (number id) – numerical - integer
  - Team - (Team Name) – categorical - string
  - g - (Games Played) – numerical - integer
  - min - (Minutes Played) – numerical - decimal
  - pts - (Points Scored) – numerical - decimal
  - reb - (Total Rebounds) – numerical - decimal
  - ast - (Total Assists) – numerical - decimal
  - stl - (Total Steals) – numerical - decimal
  - blk - (Total Blocks) – numerical - decimal
  - to - (Total Turnovers) – numerical - decimal
  - pf - (Total of Personal Fouls Committed) – numerical - decimal
  - dreb - (Total Defensive Rebounds) – numerical - decimal
  - oreb - (Total Offensive Rebounds) – numerical - decimal
  - fgm_a - (Field Goals Made - Attempted) – categorical - string
  - pct - (Field Goal Percentage) – numerical - decimal
  - 3gm_a - (3 Point Field Goals Made - Attempted) – categorical - string
  - pct_2 - (3 Point Field Goal Percentage) – numerical - decimal
  - ftm_a - (Free Throws Made - Attempted) – categorical - string

- pct_3 - (Free Throw Percentage) – numerical - decimal
- eff - (Team Efficiency) - numerical - decimal
- deff - (Team Defense Efficiency) – numerical - decimal
- year - (Years of Current NBA season) – categorical - string

VISUALIZATIONS/QUESTIONS:

- QUESTIONS: Below list 5 initial questions that could be asked about your data and what visualization you would use to answer the question. You must include a question for each of the graphs below.
    - LINE GRAPH QUESTION:
        - What is the correlation between points scored and rebounds? How has it changed from past seasons to the present?
    - HISTOGRAM QUESTION:
        - How has the number of points scored by each team changed with each season?
        - What range of points scored per game do most teams fall under?
    - BAR CHART QUESTION:
        - How does the number of assists that a team has relate to the number of 3 points made, the total points scored, and the team efficiency?
    - SCATTER PLOT QUESTION:
        - As 3gm_a have seemingly increased over recent years, does deff tend to increase or decrease?
    - Our Choice: LINE GRAPH QUESTION
        - What is the correlation between NBA Teams who have won 2+ championships in their franchise and their salaries over time?

Continued next page

- SUBGROUP COMPARISION: Include an additional question that allows you to compare information about multiple subgroups in your dataset (ideal/fair/good diamonds; males/females; Europe vs. America; etc.).
- Young vs Old
    - Does a team with a higher average of age tend to perform better or worse?
- Regular Season vs Playoffs

    - Is there an advantage of gaining a higher seed during the regular season? Do teams perform better or worse in the Playoffs?
- Salaries
    - How well does a player perform after being given a contract extension?
- Right vs Left
    - Do right handed players perform better than left handed players?

- PLAN FOR INCORPORATING A JOIN: What second table of information can you retrieve so as to use a "join" in your project? What question would this second table let you answer?

    - We will plan to use the link located below to "join" with our primary dataset throughout this project. The question we would have this answer would be our fifth question asking about the correlation between NBA Teams who have won 2+ championships in their franchise and their salaries over time.
        - https://data.world/datadavis/nba-salaries