In [1]:
```
# libraries
import pandas as pd
import numpy as np
import altair as alt
import random
```

PSTAT 100 Project plan report

This is a guide to preparing your project plan. It functions both as a guide to the work you'll need to do and as a guide to preparing the deliverable. You can use it as a template to draft the plan report; if so, please remove the text explanations of each section.

While you may find it useful initially to follow the outline given, you do not need to adhere to it exactly -- you're free to organize your submission in the way that seems most natural to you. However, please do keep the high-level sections, so that your report includes the following headers:

1. Background
2. Data description
3. Initial exporations
4. Planned work

Your report does not need to be long. It should be about 2-4 pages, and may not be much longer than this template once you replace the guiding text with your own work.

Group information

Group members:
Gabriel Fei
Heejae Park
Kristian Abad

Contributions:

1. Member 1 studied the data documentation and prepared the data description.
2. Member 2 worked on tidying the dataset and exploratory analysis.
3. Member 3 worked on tidying the dataset and exploratory analysis.
4. All 3 worked on proposed questions and approaches.

---

0. Background

This section should introduce your reader to the general topic you're engaging with in your project and explain any specialized knowledge that they may need to understand your dataset and why it's interesting. It doesn't need to be long, but should touch on the following points:

- Introduce the topic of your project.
- What area or areas of study are you in dialogue with for your project?
- What is your data about, broadly?
- What is the motivation for collecting the kind of data you're working with, and what sorts of things could you potentially learn?

You can look to the background sections in the homework assignments for examples. (There you can also see how to include images in your notebook.) The background sections of the homeworks are usually short and focused paragraphs intended to orient you to what you'll do in the assignment. They don't go into a lot of detail -- just enough to (hopefully) convince you that the data are interesting and explain any terminology or general information you may not know.

You may find it useful to write up the data description first, think about what the reader should know before they peek at your dataset, and then come back to the background section. I often write the background sections of your assignments last, once I have a sense of what kind of information would be most useful going into the assignment.

Background: Sales of videogames

Videogames have been an increasingly popular form of entertainment coming out of the crash of 1983 to the present pushing computational hardware, content creation, competitive play, and opening many commercial opportunities along the way. Titles or the games themselves have existed through various means such as consoles or specifically curated hardware from companies such as Sony with their Playstation series of console, Microsoft with Xbox consoles, and Nintendo with their most recent Switch console. Another popular option in terms of hardware comes through custom built computers.

The driving force to gather the data on already released video game titles is to understand the industry as a whole and what direction it may take. The data to be explored will lean more towards the business side of the industry, exploring the sales of titles which can be used to discover overall change in tastes over time, the longevity of a franchise or series of games, and possible differences in success of a title by different parts of the world (North America, Europe, Japan, etc). All of this and more can be leveraged could be useful to better inform indie developers and well established publishers alike to innovate through new games or potentially help evaluate how lucrative a current project may be.

---

1. Data description

This section should introduce your dataset in detail. It should reflect your having gone through the collect/acquaint/tidy stages of the lifecycle. Below I've provided you with an outline. You do not need to adhere to this strictly -- in fact, it would be more natural to divide the items among a few short paragraphs -- but you should touch on each item in a format that suits your project.

Basic information

Help your reader understand what your data is, where it came from, and how it can be used. Provide the following.

General description: provide a one- or two-sentence description of the data right at the beginning. For instance, "The data are diatom counts sampled from evenly-spaced depths in a sediment core from the gulf of California." Nothing too complicated, just something to give your reader a sense of the 'what' right off the bat.

Source: indicate where your data came from. Provide a verbal description -- who collected it as part of what project and where -- and either a citation or a hyperlink.

Collection methods: How were the data values obtained? Provide a simple description of how measurements were taken (using scientific equipment? web scraping? surveys?).

Sampling design and scope of inference: Indicate the relevant population. If identifiable from data documentation, state the sampling frame and sampling mechanism and indicate the scope of inference. If no information is available about the sampling design, indicate this instead, and discuss the extent to which having no scope of inference is a limitation for the particular topic you're investigating.

The data are recorded number of sales in games from year 1980 through 2020 that had sold more than 100,000 copies. The data can be found on https://www.kaggle.com/arslanali4343/sales-of-video-games. The intention for the collection of the data was not specified by the user who collected the data. In relevance to the source of data, it is known that the data values was web scraped off the website https://www.vgchartz.com/ using the python library Beautiful Soup, which is known to pull data from HTML or XML documents. The relevant population is the popular games introduced to the game market since 1980 and the sampling mechanism was specifically framed around games that totalled up to more than 100,000 copies sold. Additionally, it is substantial to claim that the scope of inference can be extended to games leaning more towards consoles (data is limited in that it does not contain the entirety of the Steam and other platforms) that are currently out for sale in the game market.

Data semantics and structure

Units and observations: State the observational units.

Variable descriptions: Provide a table of variable descriptions. If your dataset is large and you'll only work with a subset of the total available variables, limit your attention to the variables that you'll work with. Here's a template you can work with:

Name	Variable description	Type	Units of measurement
Rank	Ranking of Overall Sales	Numeric	Rank Numbers
Name	Name of The Game	String	Game Names
Platform	Platform of Game Release	String	Game Platforms
Year	Year of Game Release	Numeric	Calendar Year
Genre	Genre of The Game	String	Game Genres
Publisher	Publisher of The Game	String	Game Companies
NA_Sales	Total Sales in North America	Numeric	in Millions
EU_Sales	Total Sales in Europe	Numeric	in Millions
JP_Sales	Total Sales in Japan	Numeric	in Millions
Other_Sales	Total Sales in Rest of the World	Numeric	in Millions
Global_Sales	Total Sales Worldwide	Numeric	in Millions

Example rows: Print a few example rows of your dataset in tidy format. Please don't include the codes you used to manipulate the raw data. Do that in a separate notebook and export the result to a .csv file -- `data.to_csv('tidy-data.csv')` -- to load directly into the cell below.

Start your draft here.

In [2]:
```
# Load tidied data and print rows
# Ideally going to keep these rows
vg_sales = pd.read_csv('vgsales.csv')
vg_sales.head()

# Just some stuff for the next section
vg_sales.shape
# vg_sales.isna().any()

# publisher_missing = vg_sales.Publisher.isna()
# year_missing = vg_sales.Year.isna()
# global_sales_missing = vg_sales.Global_Sales.isna()

# Create missing publisher indicator variable and view true observations
# vg_sales.loc[:,'pub_missing'] = publisher_missing
# vg_sales[vg_sales.pub_missing == True]

# Create missing year indicator variable and view true observations
# vg_sales.loc[:,'year_missing'] = year_missing
# vg_sales[vg_sales.year_missing == True]

# Create missing global sales indicator variable and view true observations
# vg_sales.loc[:,'global_sales_missing'] = global_sales_missing
# vg_sales[vg_sales.global_sales_missing == True]
vg_sales_mod = vg_sales.dropna(0)
vg_sales_mod
# Note: Probably could fill in missing values for publisher and global sales but year is the one with like 271 missing observations
# ...so not sure. Could fill in the data by searching all the years and inserting them but it just might take awhile
```

Out[2]:

	Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
0	1	Wii Sports	Wii	2006	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74
1	2	Super Mario Bros.	NES	1985	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24
2	3	Mario Kart Wii	Wii	2008	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82
3	4	Wii Sports Resort	Wii	2009	Sports	Nintendo	15.75	11.01	3.28	2.96	33.00
4	5	Pokemon Red/Pokemon Blue	GB	1996	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.37
...	...	...	...	...	...	...	...	...	...	...	...
16593	16596	Woody Woodpecker in Crazy Castle 5	GBA	2002	Platform	Kemco	0.01	0.00	0.00	0.00	0.01

	Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
16594	16597	Men in Black II: Alien Escape	GC	2003	Shooter	Infogrames	0.01	0.00	0.00	0.00	0.01
16595	16598	SCORE International Baja 1000: The Official Game	PS2	2008	Racing	Activision	0.00	0.00	0.00	0.00	0.01
16596	16599	Know How 2	DS	2010	Puzzle	7G//AMES	0.00	0.01	0.00	0.00	0.01
16597	16600	Spirits & Spells	GBA	2003	Platform	Wanadoo	0.01	0.00	0.00	0.00	0.01

16289 rows × 11 columns

---

2. Initial explorations

At this stage, you may spend most of your effort on the computing side tidying up the data. You're not expected to complete a thorough exploratory analysis, and if your dataset was especially messy to start with, you may not even begin your exploratory analysis by the time you prepare this report. You have the option to leave exploration for the next stage of work and simply report basic properties of the dataset, but you should at minimum address the items in the 'basic properties' section below.

Basic properties of the dataset

Help the reader get acquainted with your dataset on a simple level by identifying characteristics of the dataset and variable summaries. Some amount of code is fine here, but try to use code cells sparingly.

Dimensions: state the dimensions of the data (in tidy format, of course).

Missing values: Are there missing values? If so, why are they missing?

Variable summaries: Provide simple variable summaries for the most important variables in your dataset. Preferably, you'll do this for all variables, but if you have a large number, you might need to prioritize and focus on the ones most of interest. What exactly you do is a little case-specific, but think of things like means and variances, min/max, number of levels and observation counts for categorical variables, etc.

Dimensions: (16598, 11)

Missing values: Year, Publisher, and Global Sales contain missing values. The missing values appear to be a result of an error in the web scraping script and data compilation process as checking some observations from the website that was scraped yields existing values. A few observations were checked but couldn't check a fair amount because of a slow and faulty website.

Variable summaries: The most important variables are the year, genre and global sales. The year has the unit measure of time and contains the time period from 1980 to 2020. Additionally, the genre consists of 14 unique genres, with examples such as Sports, Platform, Racing, Misc, etc... . Lastly, the global sales variable records the count of copies in millions that the game has sold since its release year.

Exploratory analysis

If you were lucky and your dataset was neat, you should aim to include a few exploratory plots or tables here -- they don't need to be polished at this stage, but you should select plots that are informative (rather than including all plots you may have looked at).

If you do include exploratory graphics or tables, please explain in a sentence or two what each one shows. Try to include a minimum of code. Consider saving your plots as images and inputting images into markdown cells instead of generating them anew via code cells.

In [3]:
```
vg_sales.head(20)
```

Out[3]:

	Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
0	1	Wii Sports	Wii	2006	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74
1	2	Super Mario Bros.	NES	1985	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24
2	3	Mario Kart Wii	Wii	2008	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82
3	4	Wii Sports Resort	Wii	2009	Sports	Nintendo	15.75	11.01	3.28	2.96	33.00
4	5	Pokemon Red/Pokemon Blue	GB	1996	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.37
5	6	Tetris	GB	1989	Puzzle	Nintendo	23.20	2.26	4.22	0.58	30.26
6	7	New Super Mario Bros.	DS	2006	Platform	Nintendo	11.38	9.23	6.50	2.90	30.01
7	8	Wii Play	Wii	2006	Misc	Nintendo	14.03	9.20	2.93	2.85	29.02
8	9	New Super Mario Bros. Wii	Wii	2009	Platform	Nintendo	14.59	7.06	4.70	2.26	28.62
9	10	Duck Hunt	NES	1984	Shooter	Nintendo	26.93	0.63	0.28	0.47	28.31
10	11	Nintendogs	DS	2005	Simulation	Nintendo	9.07	11.00	1.93	2.75	24.76
11	12	Mario Kart DS	DS	2005	Racing	Nintendo	9.81	7.57	4.13	1.92	23.42
12	13	Pokemon Gold/Pokemon Silver	GB	1999	Role-Playing	Nintendo	9.00	6.18	7.20	0.71	23.10
13	14	Wii Fit	Wii	2007	Sports	Nintendo	8.94	8.03	3.60	2.15	22.72
14	15	Wii Fit Plus	Wii	2009	Sports	Nintendo	9.09	8.59	2.53	1.79	22.00
15	16	Kinect Adventures!	X360	2010	Misc	Microsoft Game Studios	14.97	4.94	0.24	1.67	21.82
16	17	Grand Theft Auto V	PS3	2013	Action	Take-Two Interactive	7.01	9.27	0.97	4.14	21.40
17	18	Grand Theft Auto: San Andreas	PS2	2004	Action	Take-Two Interactive	9.43	0.40	0.41	10.57	20.81
18	19	Super Mario World	SNES	1990	Platform	Nintendo	12.78	3.75	3.54	0.55	20.61
19	20	Brain Age: Train Your Brain in Minutes a Day	DS	2005	Misc	Nintendo	4.75	9.26	4.16	2.05	20.22

In [4]:
```
# VG sales total by year
```

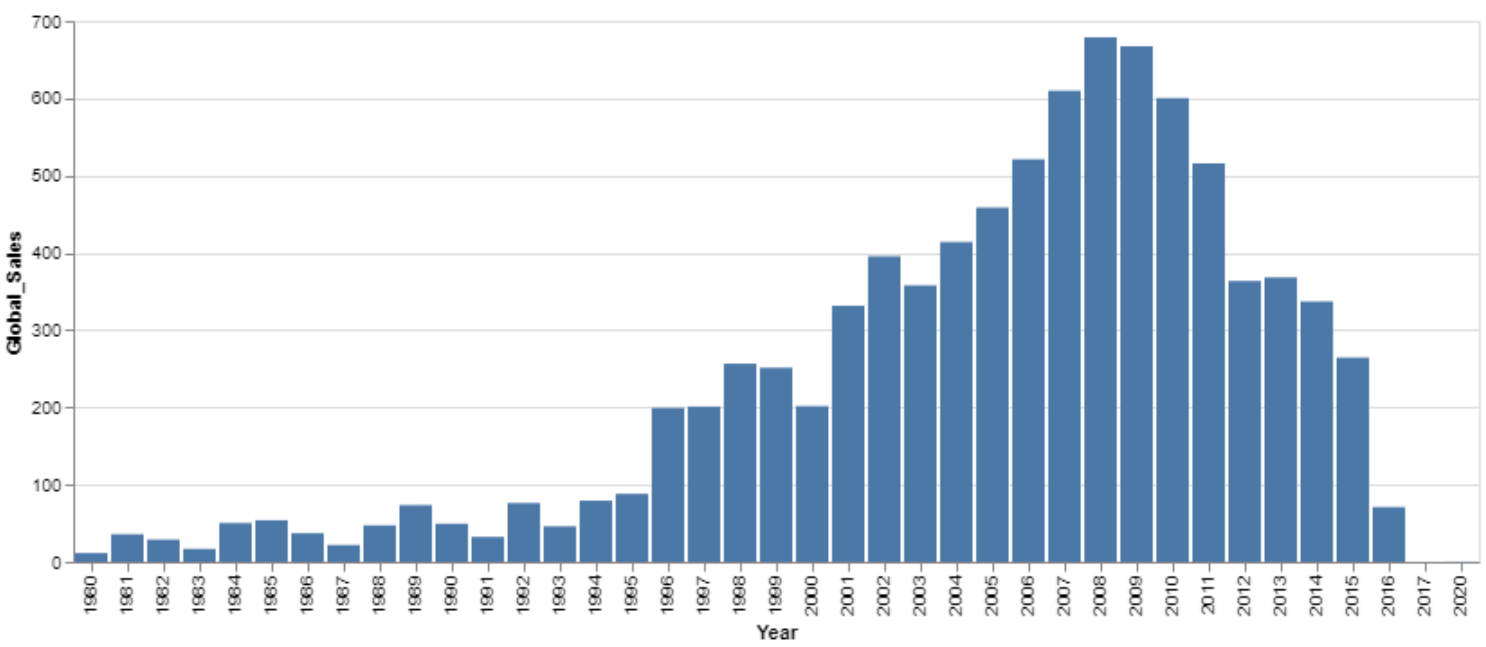
```
vg_sales_mod2 = vg_sales_mod.groupby('Year').sum().reset_index()
vg_sales_mod2
vg_sales_mod3 = vg_sales_mod.groupby('Genre').sum().reset_index().sort_values(by = 'Global_Sales')
vg_sales_mod3
# vg_sales_mod4 = vg_sales_mod.groupby(['Year', 'Genre']).count().drop(columns = ['Name', 'Platform', 'Publisher', 'NA_Sales',
#                                                                                 'EU_Sales', 'JP_Sales', 'Other_Sales', 'Global_Sales']).reset_in
# vg_sales_mod4 = vg_sales_mod4.fillna(0).reset_index()
# vg_sales_mod4
```

Out[4]:

	Genre	Rank	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
11	Strategy	6748201	67.83	43.97	49.10	11.23	173.27
1	Adventure	14679183	101.66	62.86	51.87	16.70	234.47
5	Puzzle	5496785	120.93	49.57	56.68	12.47	242.21
9	Simulation	7269349	179.99	112.25	63.54	31.36	389.98
2	Fighting	6371780	219.54	99.80	87.15	36.19	444.05
6	Racing	9699328	356.23	235.89	56.82	76.68	726.76
3	Misc	14445141	395.12	211.27	106.67	73.92	789.87
4	Platform	6019939	444.53	199.39	130.65	51.51	829.13
7	Role-Playing	11840252	325.63	186.69	350.29	59.38	923.83
8	Shooter	9399409	574.21	309.30	38.18	101.90	1026.20
10	Sports	17105195	667.66	370.90	134.99	132.65	1309.24
0	Action	25955792	859.39	513.16	158.65	184.92	1722.84

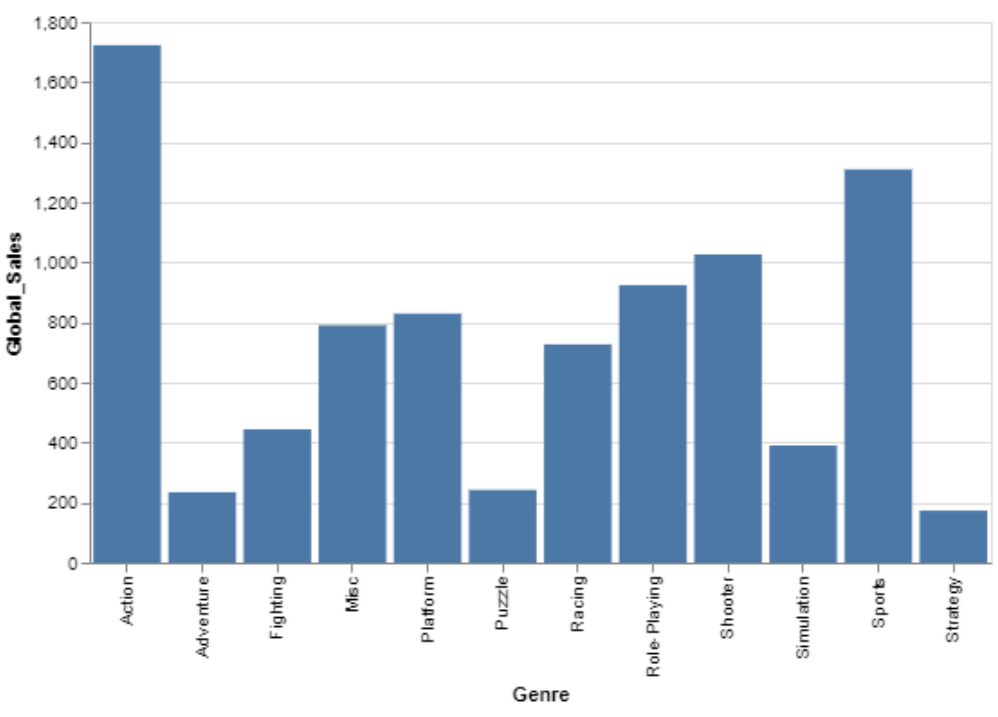
In [5]:
```
alt.data_transformers.disable_max_rows()
alt.Chart(vg_sales_mod2).mark_bar().encode(
    x=alt.X('Year'),
    y=alt.Y('Global_Sales') #sum sales
)
```

Out[5]:



In [6]:
```
alt.data_transformers.disable_max_rows()
alt.Chart(vg_sales_mod3).mark_bar().properties(
    width=500
).encode(
    x=alt.X('Genre'),
    y=alt.Y('Global_Sales') #sum sales
)
```

Out[6]:



In [7]:
```
# alt.data_transformers.disable_max_rows()
# alt.Chart(vg_sales_mod4).mark_bar().properties(
#     width=500
# ).encode(
#     x=alt.X('Year'),
#     y=alt.Y('Genre:N') #sum sales
# )
# Possible graph depicting number of titles per genre per year.
```

---

3. Planned work

Here you should indicate your tentative ideas for your analysis. Don't worry, these aren't final -- you can always change your mind later or shift gears if they don't pan out. The objective is to have you start thinking ahead about what you'll do.

Questions

Please propose two focused questions that you plan to explore.

1. How has the emergence of competing platforms affected the videogame market?
2. How have different publishers catered to specific genres over time?

Proposed approaches

For each question, please describe an idea or two about how you might approach the question.

1. We would look at the release dates of different consoles. Taking those dates and consoles(platforms) we would look at the total sales by platform to determine the success of the competing platforms before and after these dates.
2. We would look at the number of titles by genre released by different publishers each year. Then looking at sales we could determine the extent to which publishers cater to specific genres based of their success.

Submission Checklist

1. Save file to confirm all changes are on disk
2. Run *Kernel* > *Restart & Run All* to execute all code from top to bottom
3. Save file again to write any new output to disk
4. Select *File* > *Download as* > *HTML*.
5. Open in Google Chrome and print to PDF on A3 paper in portrait orientation.
6. Submit to Gradescope