PROGRAMMING IN PYTHON FOR DATA SCIENCE(CS3PP19)
Final Exam: Question 1

52944

April 27, 2021

1 Initialization Code

```
import pandas as pd
from pandas.plotting import parallel_coordinates
data=pd.read_csv("data.csv")
data=data.dropna()
```

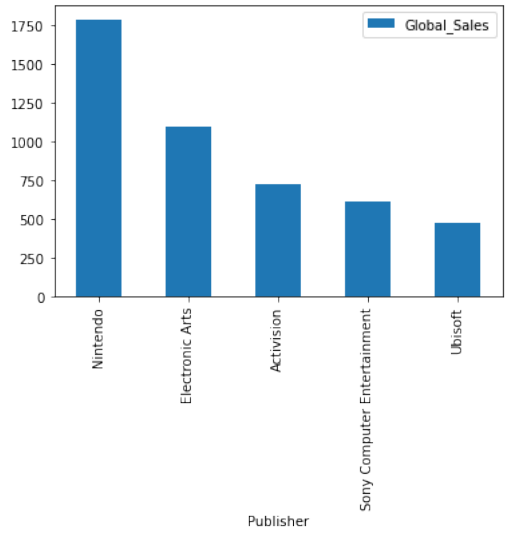
2 Q1.a.i

```
def top_publishers(df_input,att,top_num):
        return df_input[[att,"Publisher"]]\
                .groupby("Publisher")\
                .sum()\
                .sort_values(att,ascending=False)[:top_num]
```

3 Q1.a.ii

```
top5=top_publishers(data,"Global_Sales",5)
top5.plot.bar()
```

Figure 1: Global sales trend amongst 5 top publishers

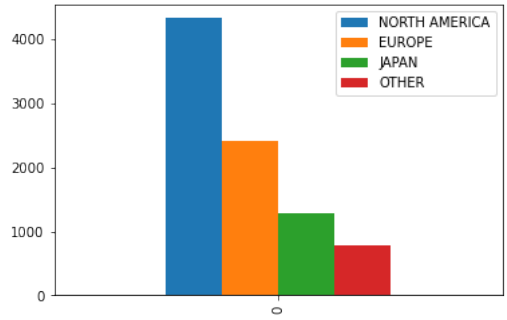


4 Q1.a.iii

4.1 Process the data to show in one graph the total sales pergeographic area (e.g. North America, Europe, etc.).

```
pd.DataFrame({
        "NORTH_AMERICA":[data.NA_Sales.sum()],
        "EUROPE":[data.EU_Sales.sum()],
        "JAPAN":[data.JP_Sales.sum()],
        "OTHER":[data.Other_Sales.sum()]}).plot.bar()
```

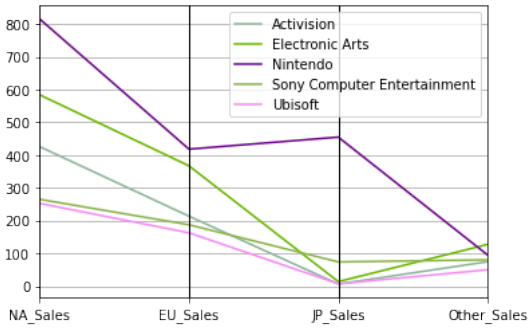
Figure 2: Global sales trend amongst 5 top publishers



4.2 How is the sales distribution in these markets divided considering the top 5 publishers in global sales?

```
sales_data=data.loc[
        (data["Publisher"].isin(list(top5.reset_index()["Publisher"])))]\
        .groupby("Publisher")\
        .sum()
parallel_coordinates(\
        sales_data[["NA_Sales","EU_Sales","JP_Sales","Other_Sales"]]\
        .reset_index(),"Publisher")
```

Figure 3: sales per region amongst 5 top publishers



4.3 Q1.b

Our strategy of curating the dataset before starting a data science project relies heavily on the project's requirements. As there are no requirements given, I will briefly explain the standard steps that need to be taken. Finally, I will demonstrate some use cases and provide steps for those.

4.3.1 Standard Preproccessing

There are several steps considered as a standard before any data science algorithm is applied to the dataset. These are ...

- Dropping the rows containing missing values with .dropna() (dataset is large enough so dropping columns is not an issue)
- Year column type should be converted to Integer (for obvious reasons)
- Global sales column should be removed, as is highly correlated with the NA,EU,JN,OTHER sales columns(summation)

4.3.2 Requirement-specific proccess

- Given more requirements, domain-specific knowledge may allow us to discard unnecessary columns. For example, in our questions (Q1.a), we perform some EDA around sales. In the specific EDA, the columns Rank, Platform, Year, Gendre could be discarded.
- Do we want to find relationships between variables or just summary statistics and complex queries? If our data science project aims to find relationships between different variables, then a normalization-outliers excluding strategy is necessary. In our EDA on question A, such a move was not needed.