Machine learning for social Science

HW 1: Python

Michelle Zhuang mz3067

Instructions: Please submit your answers as 2 files uploaded to Courseworks: a Jupyter Notebook (.ipynb) file & a pdf export. Please double check that all pages exported properly, sometimes they get cut off! In answering each of the following questions please include (a) the question as a markdown header in your Jupyter notebook, (b) the raw code that you used to generate any results, tables, or figures, and (c) the top ten or fewer rows of the dataframe (do not include more than ten rows for any table in your report). Include any plots or figures generated from your code as well.

Part A

Q1

Find the url for the mtcars dataset from the following website: https://vincentarelbundock.github.io/Rdatasets/datasets.html. Read through the "DOC" file to understand the variables in the dataset, then use the following url to import the data using pandas read_csv function.

```
# package import
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

```
# import the url of the csv
url = 'https://vincentarelbundock.github.io/Rdatasets/csv/datasets/mtcars.csv'
df = pd.read_csv(url)
```

Q2

Display the first five rows of the data.

```
# first five rows
df.head()
```

	rownames	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
0	Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
1	Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
2	Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
3	Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
4	Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2

后续步骤: 使用 df生成代码 查看推荐的图表 New interactive sheet

Q3

For each category in the cyl column, calculate the average mpg for all cars with that cyl value.

```
# use groupby, calculate the average mpg for each category in the cyl column
average_mpg_in_cyl = df.groupby('cyl')['mpg'].mean()
average_mpg_in_cyl
```

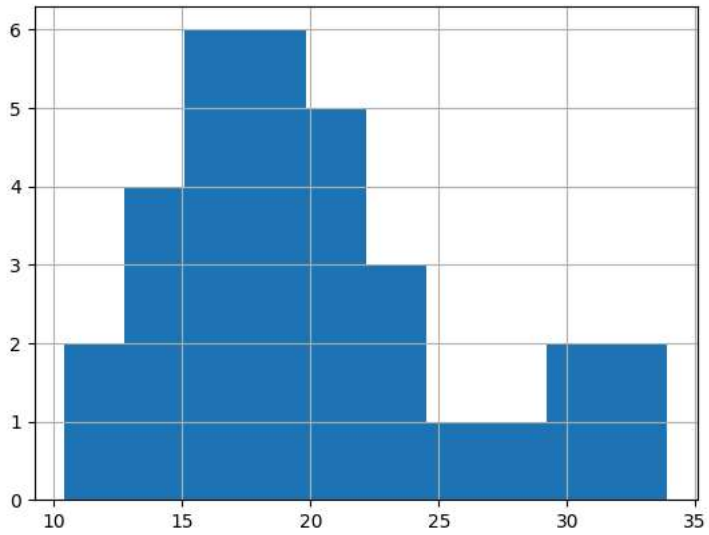
cyl	mpg
4	26.663636
6	19.742857
8	15.100000

dtype: float64

Q4

Create a histogram using the mpg column.

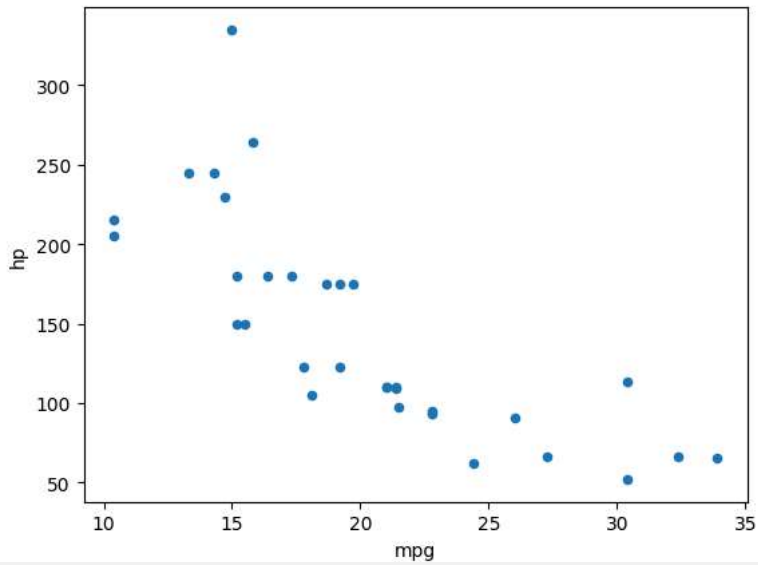
```
# create a histogram of mpg column
df['mpg'].hist()
plt.show()
```



Q5

Choose two variables in the data and create a scatterplot.

```
# use mpg and hp to create scatter plot
df.plot.scatter(x='mpg', y='hp')
plt.show()
```



Part B

Find a tabular dataset that interests you that has "tidy" data.

- Tidy data has data that is ready for your data analysis.
- For our tasks we want something where columns represent variables and rows represent unique observations.
- Give a brief description of the dataset. Provide a citation of the dataset (any format is fine).

Dataset description

Data Name: Video Game Sales
Data link: https://www.kaggle.com/datasets/gregorut/videogamesales
This data contains Variables

- Name: The title of the video game.
- Platform: The gaming platform (e.g., PS4, Xbox One, PC).
- Year: The year the game was released.
- Gerne: The genre of the game (e.g., Action, Sports, RPG).
- Publisher: The company that published the game
- Different region sales: variable such as *NA_Sales, EU_Sales, JP_Sales, Other_Sales, Global_Sales*

```
# import the csv file
file = 'https://raw.githubusercontent.com/michellezzmmmm/interview/main/vgsales.csv'
vgame_df = pd.read_csv(file)
```

Q1

Display the first five rows of the data.

```
vgame_df.head()
```

	Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
0	1	Wii Sports	Wii	2006.0	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74
1	2	Super Mario Bros.	NES	1985.0	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24
2	3	Mario Kart Wii	Wii	2008.0	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82
3	4	Wii Sports Resort	Wii	2009.0	Sports	Nintendo	15.75	11.01	3.28	2.96	33.00
4	5	Pokemon	GB	1996.0	Role-	Nintendo	11.27	8.89	10.22	1.00	31.37

后续步骤： 使用 vgame_df生成代码 查看推荐的图表 New interactive sheet

Q2

Create a visualization using one or two variables from this data.

```
# I want to select the top 10 games by the global salse and plotting the sales
top_10_games = vgame_df.sort_values(by='Global_Sales', ascending=False).head(10)
top_10_games.plot.bar(x='Name', y='Global_Sales')
plt.xlabel('Name')
plt.ylabel('Global Sales')
plt.title('Top 10 Games by Global Sales')
plt.show()
```