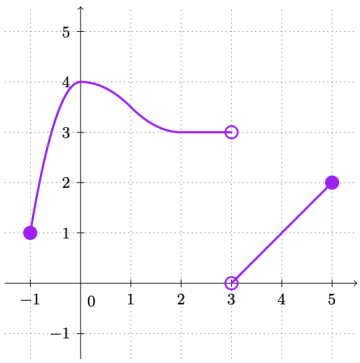
Mathematical Foundations of Data Science
Assignment 1

Trimester 1, 2023

Chang Dong (a1807402)

¶

1. Consider the function g whose graph is below



(a) What is the domain of g?

Answer:

The domain of g is $\{x \in R \mid -1 \leqq x < 3,\ 3 < x \leqq 5\}$.

(b) What is the range of g?

Answer:

The domain of g is $\{y \in R \mid 0 < x \leqq 4\}$.

(c) Is g a one-to-one function? Explain your answer.

Answer:

No, because when $y \in [1, 2]$, it conresponds to two x values.

2. In total, how much money do the people in Adelaide spend buying cups of coffee every year?
Note that you are not expected to get an exact answer! This is an exercise in *estimation* - there is no one correct number that will get you full marks. Rather, you must give a reasonable argument with plausible estimates following the principles presented in the course materials. You need to justify your assumptions and make it clear how you arrived at your answer.

Answer:

The price of one cup of coffee is around \$5. 1 year equals around 50 weeks, and each week has 5 weekdays, so the total weekdays are around 250 days. Besides, there are around 10 public holidays, and people have legal anual leave which are around 30 days, so total holidays are around 50 days. Thus, total days people need to work are around 200 days. And I suppose that everyone will cosume 1 cup of coffee every working day.

So the total cost of one person is around $\$5 \times 200 = \$1,000$.

As for the the total cost of all the people in Adelaide. There are around 1M people live in adelaide, I suppose that everyone comsumes coffee include little children and unemployed adult though it seems not correctly enough.

So the total cost of all the people in Adelaide is around $\$1,000 \times 1,000,000 = \$1\ Billion$

3. Let $J = \{x \in N \mid -7 \leq 3x - 4 < 9\}$, $K = (0, 3]$, and $L = \{0, 1/2, 3\}$.

Determine the following, giving reasoning:

(a) $J \cap K$
Answer:
$J = \{x \in N | -7 \leq 3x - 4 < 9\} = \{x \in N | -1 \leq x < 4\} = \{0, 1, 2, 3\}$
$K = (0, 3]$
Given that, $J \cap K$ means the intersection between J set and K set, which is $\{1, 2, 3\}$

(b) $K \setminus L$
Answer:
$K = (0, 3]$
$L = \{0, 1/2, 3\}$
Given that, $K \setminus L$ means the elements in the part of K that not included in L, which is $\{x \in R | 0 < x < 1/2, 1/2 < x < 3\}$

4. You should complete this question using a Jupyter Notebook. All of the code you will need to complete this question can be taken directly or generalised from the week 1 computer exercise, or will be given to you in the question.

Download the file movie data.csv from MyUni (this file comes from the imdb-5000-movie-dataset on Kaggle). Then use Python to do the following:

(a) Using pandas, read the data into a dataframe and print out its tail().

In [1]:
```
import pandas as pd

df = pd.read_csv("movie_data.csv")
df.tail()
```

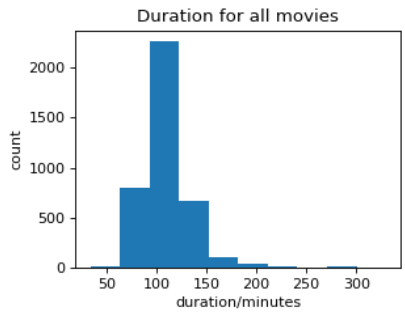
Out[1]:

	Unnamed: 0	duration	gross	genres	movie_title	num_voted_users	plot_keywords	movie_
3885	5033	77	424760	Drama|Sci-Fi|Thriller	Primer	72639	changing the future|independent film|invention...	http://www.imdb.com/title/tt
3886	5034	80	70071	Thriller	Cavite	589	jihad|mindanao|philippines|security guard|squa...	http://www.imdb.com/title/tt
3887	5035	81	2040920	Action|Crime|Drama|Romance|Thriller	El Mariachi	52055	assassin|death|guitar|gun|mariachi	http://www.imdb.com/title/tt
3888	5037	95	4584	Comedy|Drama	Newlyweds	1338	written and directed by cast member	http://www.imdb.com/title/tt
3889	5042	90	85222	Documentary	My Date with Drew	4285	actress name in title|crush|date|four word tit...	http://www.imdb.com/title/tt

(b) Create a histogram of duration for all movies.
Hint: Remember to add axis labels.

In [2]:
```
import matplotlib.pyplot as plt

plt.figure(figsize=(4, 3), dpi=80)
plt.hist(df.duration)
plt.xlabel("duration/minutes")
plt.ylabel("count")
plt.title("Duration for all movies")
plt.show()
```



(c) Calculate the mean duration of all movies. Print out a statement showing the mean duration.
Hint: In the week 1 computer exercise, we saw how to print out a line containing text and numbers.

In [3]:
```
mean_duration = df.duration.mean()
print("the mean duration of all movies is {:.1f} minutes".format(mean_duration))
```

```
the mean duration of all movies is 109.9 minutes
```

(d) Among movies longer than 180 minutes, find the movie with the highest gross.
Hint: Try creating a new dataframe that only contains movies with duration greater than 180.

In [4]:
```
df2 = df[df.duration > 180]
df_gross_max = df2[df2.gross == df2.gross.max()]
print("The movie longer than 180 minutes with the highest gross {} is \"{}\""
      .format(df_gross_max["gross"].values[0],df_gross_max["movie_title"].values[0].strip()))
```

```
The movie longer than 180 minutes with the highest gross 658672302 is "Titanic"
```

(e) Find the Western movie with the highest gross among movies longer than 180 minutes.
Hint: Many movies have multiple genres. We want to include all movies with Western in the genre, not just movies where Western is the only genre. The string method .str.contains() might be useful here.

In [5]:
```
df3 = df[df.genres.str.contains("Western")][df.duration > 180]
df3_gross_max = df3[df3.gross == df3.gross.max()]
print("The western movie longer than 180 minutes with the highest gross {} is \"{}\""
      .format(df3_gross_max.gross.values[0],df3_gross_max.movie_title.values[0].strip()))
```

```
The western movie longer than 180 minutes with the highest gross 184208848 is "Dances with Wolves"
```

```
/var/folders/s6/hxvlbvjs7sj27chtbkvktf180000gn/T/ipykernel_85708/2339934168.py:1: UserWarning: Boolean Series key wi
ll be reindexed to match DataFrame index.
  df3 = df[df.genres.str.contains("Western")][df.duration > 180]
```

Present your answers as a full Jupyter Notebook. Your notebook must include code to find the results, and text answering the questions based on the output of your code. Download this notebook and convert to a PDF and submit with your assignment.

Hint for submitting: You can "Download As PDF" in Jupyter, but that may not work on your computer. If it doesn't, you can download as HTML and convert that to a PDF. Make sure you join it to your assignment to make a single PDF when submitting! You might want to try googling things like "convert html to pdf" and "combine multiple pdfs". There is also a video in the Python Module on MyUni demonstrating how to save a Jupyter Notebook as a PDF.

Type *Markdown* and LaTeX: α^2