

Project 01 Answer Key

Question 3a (2 pts)

Some airline data is located at: `/class/datamine/data/flights` Roughly how much data is stored in this directory? We do not need an exact count (yet). Just give a sense: several kilobytes? megabytes? gigabytes? petabytes?

File browser solution

- If we navigate to `/class/datamine/data/flights/` in the File Manager, it reports that the airline data is 83.1 GB.

Command line solution

*# Use the `du` command with the `-b` option for approximate disk usage
and `-h` option for human-readable output.*

```
du -bh /class/datamine/data/flights
```

```
>>>
```

```
78G /class/datamine/data/flights/
```

- The airline data is about 78 GB.

Question 3b (2 pts)

Find the map files for the taxi data (these are jpg files). How many files are there, in that maps directory?

File browser solution

- The map files for the taxi data are located in `/class/datamine/data/taxi/maps/` and, in particular, there are 5 maps in that directory.

Command line solution

*# Use `ls` to list the files in `.../taxi/maps` and count the number of lines in
the output.*

```
ls /class/datamine/data/taxi/maps | wc -l
```

```
>>>
```

```
5
```

- There are 5 files in the maps directory.

Question 3c (2 pts)

Roughly how much data is stored (altogether) in the directory about **yellow** taxi cabs?

File browser solution

- The data about yellow taxi cabs is stored in the directory `/class/datamine/data/taxi/yellow/` and the File Manager reports that there are 241.9 GB of data about yellow taxi cabs in that directory.

Command line solution

```
# Use the du command with the -b option for approximate disk usage and -h  
# option for human-readable output.
```

```
du -bh /class/datamine/data/taxi/yellow  
>>>  
226G    /class/datamine/data/taxi/yellow/
```

- The yellow taxi cab data is about 226 GB.

Question 3d (2 pts)

For which years do we have election data?

File browser solution

- As seen in the directory `/class/datamine/data/election/` we have data about the elections held in the even-numbered years from 1980 to 2020.

Command line solution

```
# Use ls to list the files in ../data/election.
```

```
ls /class/datamine/data/election  
>>>  
itcont1980.txt  itcont1992.txt  itcont2004.txt  itcont2016.txt  
itcont1982.txt  itcont1994.txt  itcont2006.txt  itcont2018.txt  
itcont1984.txt  itcont1996.txt  itcont2008.txt  itcont2020.txt  
itcont1986.txt  itcont1998.txt  itcont2010.txt  
itcont1988.txt  itcont2000.txt  itcont2012.txt  
itcont1990.txt  itcont2002.txt  itcont2014.txt
```

- The election data covers the even-numbered years from years 1980 to 2020.

Question 3e (2 pts)

For which cities in California do we have AirBnB data?

File browser solution

- In the directory `/class/datamine/data/airbnb/united-states/ca/` we have data for 8 locations: 5 of them are cities and 3 are counties. The cities are: Los Angeles, Oakland, Pacific Grove, San Diego, San Francisco.

Command line solution

```
# Use ls to list the files in ../data/airbnb/united-states/ca.  
ls /class/datamine/data/airbnb/united-states/ca
```

```
>>>
```

```
los-angeles  pacific-grove  san-francisco    santa-clara-county  
oakland      san-diego       san-mateo-county  santa-cruz-county
```

- The 5 cities are Los Angeles, Oakland, Pacific Grove, San Diego, San Francisco.