Data Description and Proposed Projects

BC Data Science Workshop -- August 21, 2017

Borhan Sanandaji

Agenda

1. **Data Description**
   - Description
   - Available Dataset
   - How to Load
2. **Proposed Projects**
   - Basic Vehicle Data Statistics
   - Multiple Drivers Detection
   - Battery Voltage Forecasting

Data Description

Data Size:

- almost every 10 seconds one message per device when running
- on average 150 messages per trip
- currently recording 1 million trips per day

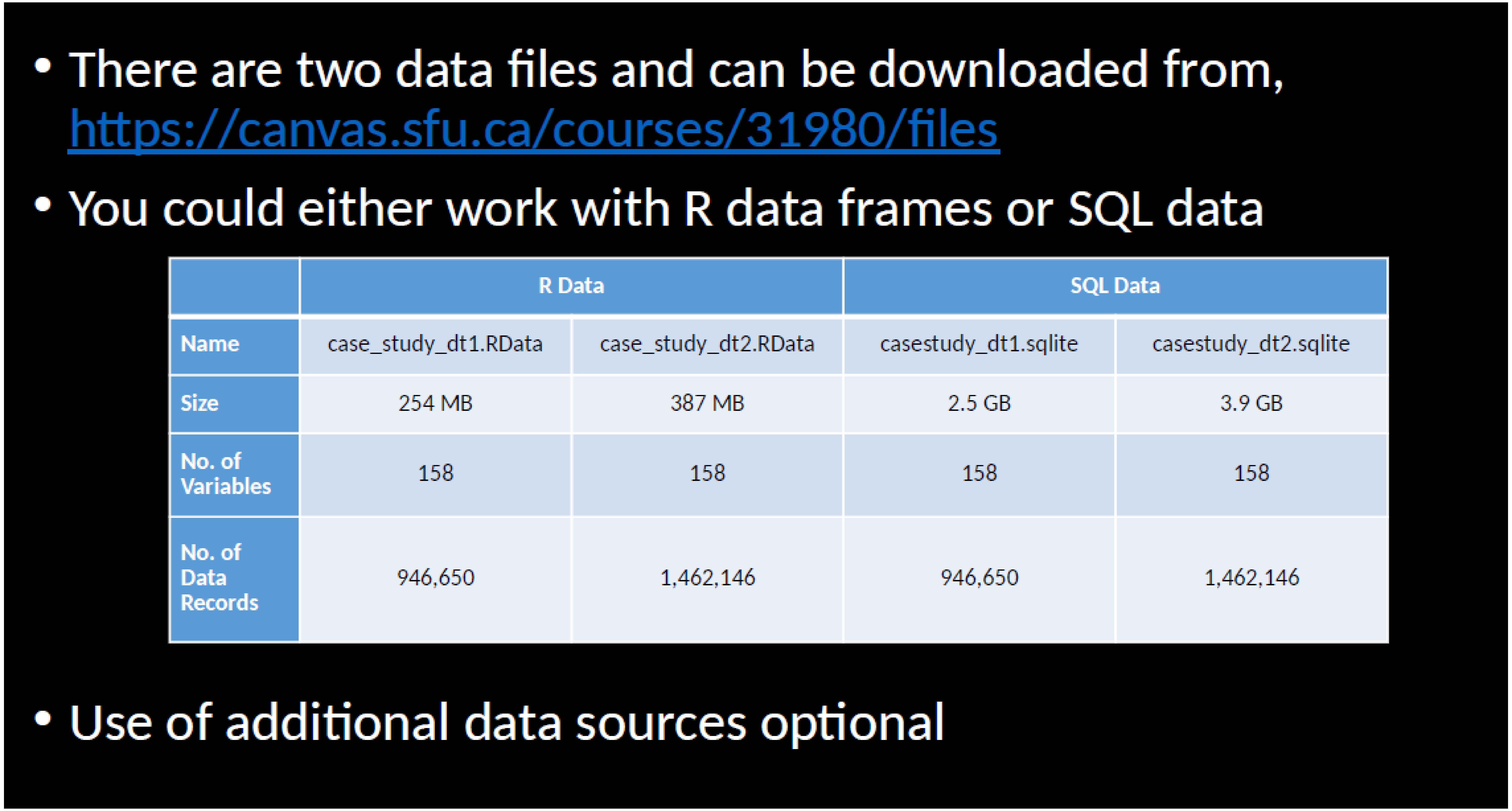
time-invariant variables:

- Vehicle Make
- Vehicle Model
- Vehicle Year
- …

time-variant variables:

- speed (times series data)
- RPM (time series data)
- Fuel Efficiency (time series data)
- …

Available Datasets

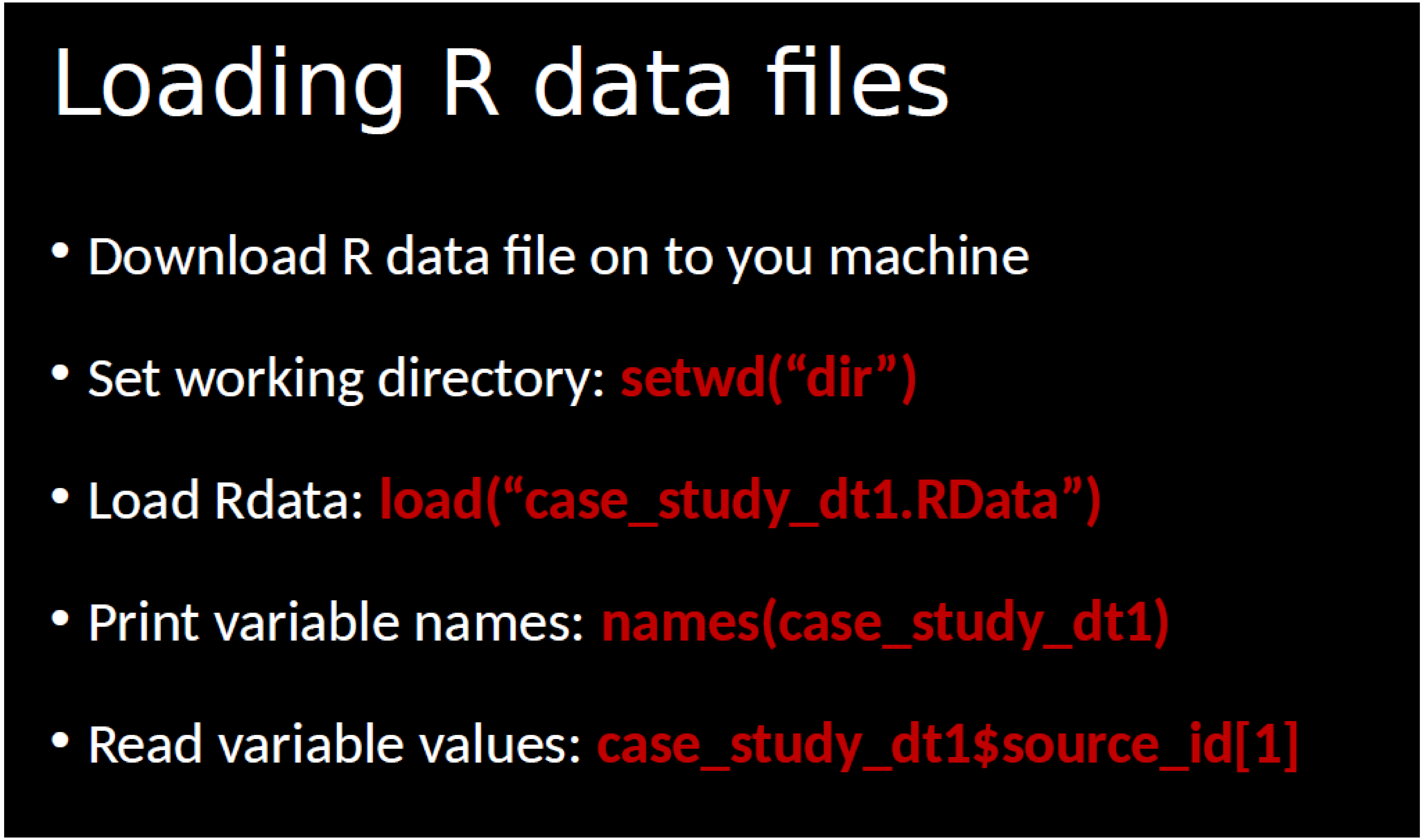


- There are two data files and can be downloaded from, https://canvas.sfu.ca/courses/31980/files
- You could either work with R data frames or SQL data

	R Data		SQL Data	
Name	case_study_dt1.RData	case_study_dt2.RData	casestudy_dt1.sqlite	casestudy_dt2.sqlite
Size	254 MB	387 MB	2.5 GB	3.9 GB
No. of Variables	158	158	158	158
No. of Data Records	946,650	1,462,146	946,650	1,462,146

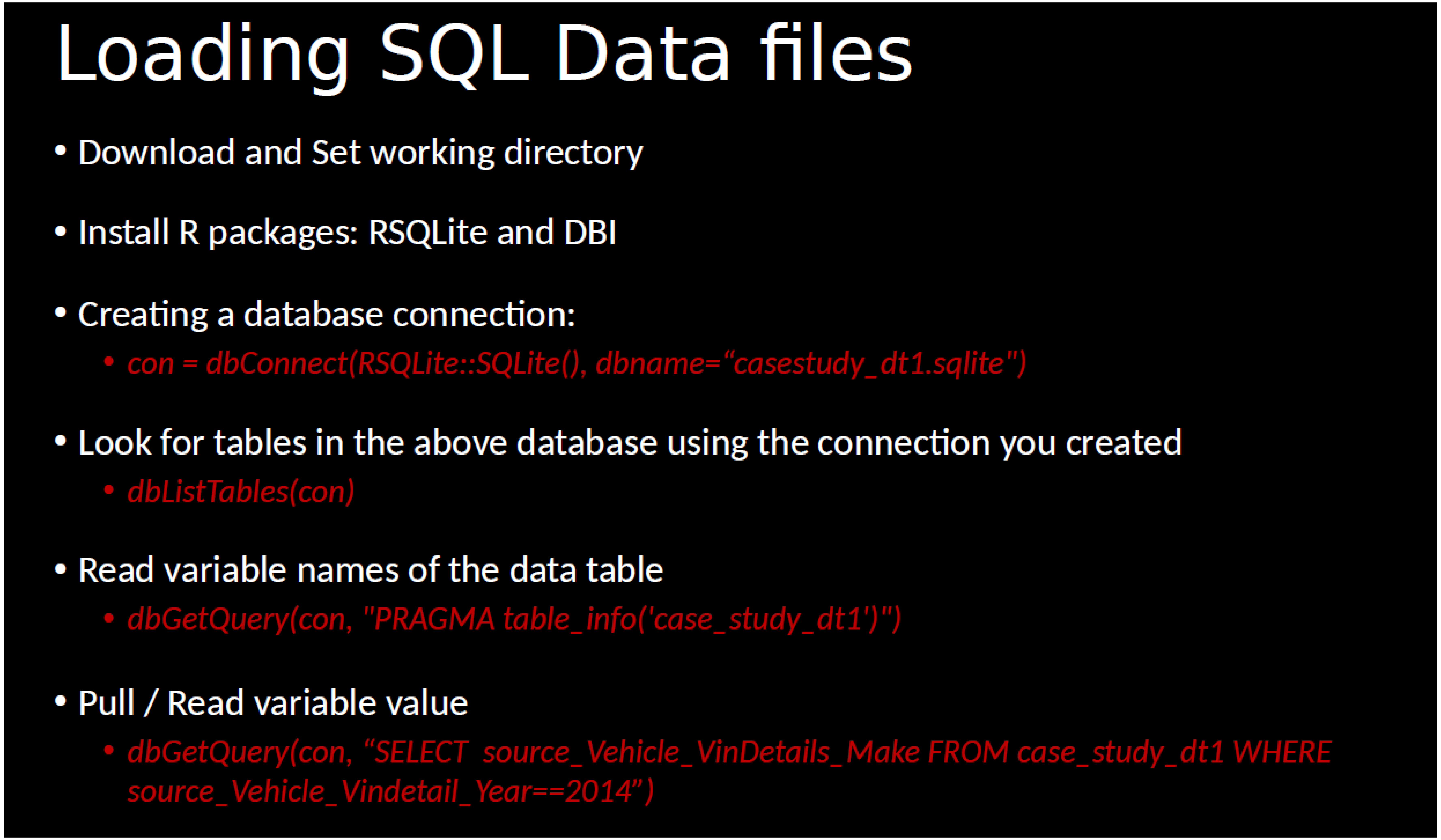
- Use of additional data sources optional

How to Load



Loading R data files

- Download R data file on to you machine
- Set working directory: **setwd(“dir”)**
- Load Rdata: **load(“case_study_dt1.RData”)**
- Print variable names: **names(case_study_dt1)**
- Read variable values: **case_study_dt1$source_id[1]**



Loading SQL Data files

- Download and Set working directory
- Install R packages: RSQLite and DBI
- Creating a database connection:
  - *con = dbConnect(RSQLite::SQLite(), dbname=“casestudy_dt1.sqlite")*
- Look for tables in the above database using the connection you created
  - *dbListTables(con)*
- Read variable names of the data table
  - *dbGetQuery(con, "PRAGMA table_info('case_study_dt1')")*
- Pull / Read variable value
  - *dbGetQuery(con, “SELECT source_Vehicle_VinDetails_Make FROM case_study_dt1 WHERE source_Vehicle_Vindetail_Year==2014”)*

Project Proposals

Idea 1:

- derive basic insights from our dataset
- possible options:
    - data visualization and summary of make-model-year distribution of vehicles
        - consider geographical distribution as well
    - design an algorithmic way to detect outliers and inaccurate values
    - find missing values, replace them with something reasonable (e.g. imputation method)
- any other interesting facts about the available dataset

Idea 2:

- multiple drivers per car detection
- “source_id” is a unique identifier of each car
- possible variables to consider:
  - “source_DeviceTime_TimeStamp” and “source_DeviceTime_Status”
  - “source_Vehicle_Location_Lat” and “source_Vehicle_Location_Lng”
  - “source_Vehicle_Speed_Value”

Idea 3:

- forecast battery voltage given the existing information in the data-set
    - ➢ start with linear regression
    - ➢ use external data resources to improve performance
    - ➢ use more advanced algorithms such as KNN

- compare results using different metrics (e.g., MAE, RMSE, etc.)
- drive results for different prediction horizon
    - ➢ minutes
    - ➢ hours
    - ➢ days
- challenges:
    - ➢ values sent when ignition off are more accurate
    - ➢ different people make different number of trips

Questions

Borhan Sanandaji
Data Scientist
borhans@moj.io

Narayan Sainaney
CTO
narayans@moj.io

Thank You!

OFFICES

Canada
1080 Howe St, 9th Floor
Vancouver, BC
V6Z 2T1

United States
4005 Miranda Ave, #100
Palo Alto
94304

CONTACT

info@moj.io
+1-855-556-6546
www.moj.io